Identification of gut microbiota markers for Inflammatory Bowel Diseases in children

Early diagnostic potentials

Felix Chinweije Nwosu

Høgskolen i **Hedmark**

Master degree in Applied Biotechnology

HEDMARK UNIVERSITY COLLEGE

Hamar, October 2011

Acknowledgement

This Masters degree thesis was carried out at the The Department of Natural Sciences and Technology, Hedmark University College, Hamar Norway, in the 2010/2011 academic year. Supervision has been provided by Professor Knut Rudi.

I would like to appreciate the excellent and enthusiastic supervision provided by my supervisor in the course of the completion of this project and for his valuable guidance in the writing process. All this I am grateful for and hope to adopt in my future endeavours.

I am equally grateful to all staff and lecturers at the department for their contribution and assistance in one way or another in my time at the college. Also, my fellows in the biotechnology group, which has made the experiences memoriable.

Finally, I would like to acknowledge my family for their prayers and best wishes all the time. Special thanks to my church family too, who has made me me feel at home in Hamar with their warm and loving heart.

Thank you Almighty God for Who my life and existence is a testimony.

Felix C. Nwosu

Hamar, 1st of October 2011.

Table of contents

Abbreviations

Ahus	Akershus University Hospital
ARISA	Automated ribosomal intergenic spacer
ASCA	anti-*Saccharomyces cerevisiae* antibody
ATG16L1	Autophagy-related protein 16-1
CD	Crohn's Disease
ddNTPs	Dideoxy nucleoside triphosphates
DGGE	Denaturing gradient gel electrophoresis
DMF	dimethyl formamide
DNA	Deoxyribonucleic acid
dsDNA	double stranded Deoxyribonucleic acid
GA	Genetic Analysis AS
GIT	Gastrointestinal tract
IBD	Inflammatory bowel diseases
IBDU	Inflammatory bowel diseases unclassified
ITS	intergenic transcribed spacer
MCR	Mutivariate Curve Resolution
NF	nuclear factor
NOD2/CARD15	nucleotide-binding oligomerisation domain 2/caspase activation and recruitment domain 15
PA	peak area
P-ANCA	Perinuclear anti-neutrophilic cytoplasmic antibody
PCA	Principal Components Analysis
PCR	polymerase chain reaction
PH	peak height
RDP	Ribosomal Database Project

rRNA	ribosomal RNA
SNE	single nucleotide extension
SNP	single-nucleotide polymorphism
SSU	small subunit
TGGE	Temperature Gradient Gel Electrophoresis
T_H	T-helper cells
TNF-α	Tumor necrosis factor-alpha
T-RFLP	Terminal Restriction Fragment Length Polymorphism
UC	Ulcerative colitis

Abstract

There are a growing body of evidences that the human gut microbiota is involved in the pathogenesis of inflammatory bowel diseases (IBD). Two mains forms exists, namely Crohn's disease (CD) and Ulcerative Coilitis (UC). Although these knowledge are so far limited, a better understanding of the gut microbes and variations in health and disease can contribute to inflammatory bowel disease diagnotics.

In this current project as reported, was designed to determine correlations between the composition of the gut microbiota and IBD in children (less than 18 years old) and thus evaluate potentials for early diagnostics. From a repertoire of faecal specimens from early diagnosed IBD (CD and UC) and non-IBD controls at Akershus University Hospital bio-bank, 75 children samples were analysed utilising bacterial small subunit (SSU) rRNA as biomarkers. DNA from this samples were extracted and purified using the QIAGEN QIAamp® DNA Stool Kit. Using a combination of forward and reverse primers that generally target the conserved regions of 16S rRNA gene of multiple microorganisms, the amplicons were direct sequenced. The mixed pool of sequences were subjected to multivariate curve resolution with resulting five components matched by Ribosomal Database Project search to, *Faecalibacterium*, *Dialister*, *Haemopilus*, Enterobacteriaceae and *Bacteroides* as dominant phylotypes and corresponding frequencies in all samples were obtained. To isolate pure bacterial species within the cohort, MCR selected samples were cloned into plasmid vectors of competent *E. coli* cells (TOPO®TA Cloning Kit) and sequenced. The clone library revealed earlier enumerated bacterial species in IBD cases against controls in literature and as well as *Haemophilus*, a previously uncharacterised gut bacteria. Similary, the clone library sequences were subjected to the Genetic Analysis (GA) in house probe design programme with specific settings that would generally label targets, excluding non-targets at the set conditions were screened on the cohort samples 16S rRNA amplicons utilising the GA two step single nucleotide extension (SNE) novel probe technology. In addition to six probes being screen, a Universal probe was screened for normalisation.

To establish correlation to disease, data generated from both multivariate resolution and probe screening was put through the SYSTAT binary logistic regression model. The MCR prediction showed *Faecalibacterium* and *Enterococcus* significantly correlating to disease,

IBD and UC respectively. However, the selective probes showed specificity to *Dialister* and *Haemophilus* clusters, correlating to UC and CD respectively and hence the diagnostic potentials proposed.

The finding in this study may have wider implication for early IBD diagnostic if followed up with extensive validation.

1. Introduction

1.1 Preamble

The human gut microflora is thought to be implicated in health and disease. (Hanauer, 2006). Increasing evidence currently exists from different studies on the effect of the gut microbiota in the development of conditions such as obesity, inflammatory bowel diseases, intestinal cancers and atopic diseases. However, our basic understanding of any underlying dynamics and variations in the gut microbiota to the individuals health and disease are very limited. Exploring these experimental lines of thought, can contribute to diagnostics, health care and lower morbidity due to chronic disease such as in idiopathic inflammatory bowel disease (IBD). (Mai & Draganov, 2009; Reiff & Kelly, 2010)

1.2 The Human gut microbiota

From parturition the human body becomes host to colonisation of a vast array of microbes. The environmentally exposed surfaces such as the skin, mouth, gut and vagina generally harbour these foreign microorganisms. (Round & Mazmanian, 2009) They develop co-dependently, consequently aiding the host with coding capacity and metabolic activities and often described with the term “superorganism” (Dethlefsen, McFall-Ngai, & Relman, 2007; Ley, Peterson, & Gordon, 2006; Mai & Draganov, 2009; Shanahan, 2009; Ventura et al., 2009). The nutrient-rich human gut bears a bulk of the diverse bacterial colonisation, an understanding that has implication in diseases and management of human health. (Dethlefsen, et al., 2007; Ley, et al., 2006)

A body of research alludes that the composition of the gut microbiota is a result of both selective pressure imposed by the host and modulated through competitions among constituent bacterial residents. (Shanahan, 2009; Ventura, et al., 2009) The human GIT being the primary site of interaction between host and bacterial species involves various associations according to roles and effects. The symbiotic interactions between the human gut and the bacterial load has been described as mutualistic, commensal or pathogenic associations. (Round & Mazmanian, 2009; Saarela, Lähteenmäki, Crittenden, Salminen, & Mattila-Sandholm, 2002; Shanahan, 2009; Ventura, et al., 2009) Within the human gastrointestinal tract (GIT), the identified microflora from culture-independent studies has

revealed well over 1000 phylotypes of about 7000 strains within 8 major phyla (Ventura, et al., 2009).

The human gut microflora has been shown to be impacted on host nutrition, immune responses and pathogenic invasion (Skånseng, Kaldhusdal, & Rudi, 2006; Ventura, et al., 2009) and eventually providing certain functions which humans cannot perform in accordance with its lineage (Freilich et al., 2010; Qin et al., 2010). To better undertand these impacts on the host, a good understanding of the microbiota may allude to these interaction. Hence, bacterial ecological community models of evolution enables predictions on the ability of microorganisms to inhabit its host. (Freilich, et al., 2010; Woese, 1987) No other is this understanding relevant than the impact the human gut microbiota exacts on the host processes which may underline it genes, cells and brains among others (Woese, 1987).

1.3 Inflammatory Bowel Disease (IBD)

Inflammatory bowel diseases are a group of conditions of inflammation affecting the intestine, colon or anywhere within the gastrointestinal tract. The incidence of IBD has been characterised by an intermittence of a set chronic inflammation of the gastrointestinal mucosa, from the mouth to the anus (*Figure* 1.3.1). (Braun & Wei, 2007; Kuehbacher et al., 2008) IBD patients exhibit symptoms like abdominal pain, emesis, diarrhea, rectal bleeding, fever, weight loss, odema, fatigue, general discomfort (malaise) and may flare up (relapse) after symptoms free periods (remission) (Braus & Elliott, 2009; Reiff & Kelly, 2010; E F Stange et al., 2006).

In addition to these symptoms, an exhibition of extra-intestinal inflammation in joints, eyes, skin, mouth and the liver could be further manifestation of IBD patients (Bouma & Strober, 2003). This could be important in diagnosis which does not only cite manifestation of intestinal symptoms (Karlinger, Györke, Makö, Mester, & Tarján, 2000).

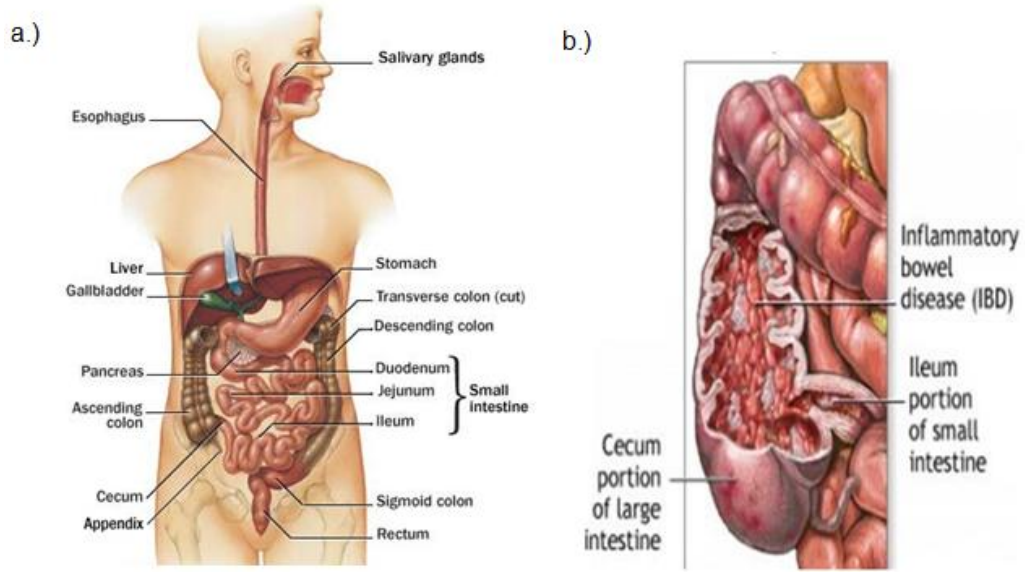

***Figure* 1.3.1**: Anatomy of the GIT (a) and Illustrating inflammation of the bowel (b). IBD characterised by inflammation could in theory affect any portion of GIT. [Image Sources: http://www.nlm.nih.gov/medlineplus/ency/imagepages/19219.htm and http://www.riversideonline.com/health_reference/Test-Procedure/MY00138.cfm]

The main forms of these set of diseases are Crohn's disease (CD) and Ulcerative colitis (UC). They both often exhibit similar characteristics features, however, it is distingushable clinically in the localisation of histopathological inflammation by screening and endoscopic features (*Figure* 1.3.2) and as well as through a profile of familial aggregation. Typically, diffused mucosal inflammation in UC cases are manifested retrogradingly in the rectum (large bowel) as pseudopolyps and may extend the entire colon in severe cases while CD involves skip lesions anywhere within the GIT (upper region) but most common in the terminal ileum (*Figure* 1.3.1). (Braun & Wei, 2007; Braus & Elliott, 2009; Salzman & Bevins, 2008; Sokol, Lay, Seksik, & Tannock, 2008; Xavier & Podolsky, 2007) Each diagnosed case can also be graded according to extent and severity of diseases, such as the number of blood stools per day (Baumgart & Sandborn, 2007; E F Stange, et al., 2006; E. F. Stange et al., 2008).

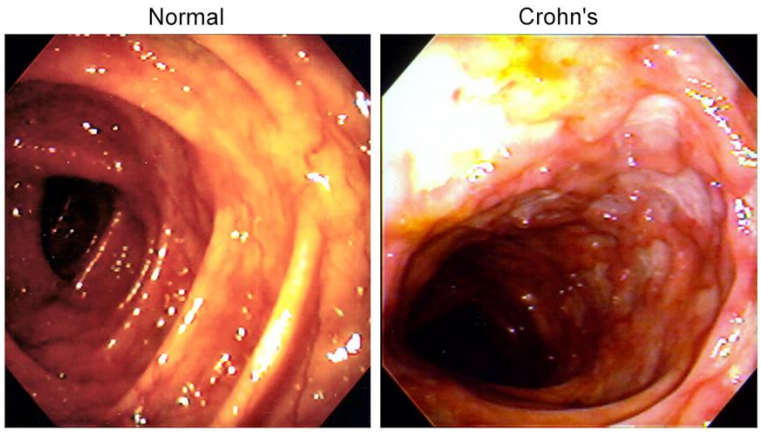

***Figure* 1.3.2**: A colonoscopic image illustrating a healthy (Normal) and diseased (Crohn's) of the transverse colon. Diseased colon shows numerous and deep ulceration of colonic mucosa. (Baumgart & Sandborn, 2007)

1.3.1 Epidemiology

There remains epidemiological suggestions that IBD prevalence and risk occurs mostly in the industralised nations of the West; North America and Europe (Braun & Wei, 2007; Braus & Elliott, 2009) within a North-South slope (E F Stange, et al., 2006; E. F. Stange, et al., 2008). Similarly, current studies are revealing a constant rise in the cases of IBD in the developing worlds especially in the industrially emerged Asia (Braun & Wei, 2007; Braus & Elliott, 2009; Xavier & Podolsky, 2007). Nevertheless, most patients tend to live normal productive lifes (Hanauer, 2006).

The onset of IBD tend to take place within the second and third decades of life (Xavier & Podolsky, 2007). Peaking firstly between 15 to 30 years, gradual decrease to remission and then slightly relapse in advanced ages, 50 to 70 years (Hanauer, 2006). However, about 5 years ago the prevalence of IBD patients in the USA and Europe are estimated at 1.4 and 2.2 million respectively (Braun & Wei, 2007; Braus & Elliott, 2009). There are generally no significant differences in IBD incidence in both male and females but there are slight inclination to CD in females (from 1:1 to 1.8:1) and UC in males. (Hanauer, 2006; Karlinger, et al., 2000)

1.3.2 Etiology

Currently, the precise aetiology of IBD has not been completely elucidated. However, the pathophysiological observation from various experimental studies and also involving animal models, suggest a multifactorial interplay involving the enteric microbiota, host genetic factor and immune function and with environmental implications (*Figure* 1.3.3). (Braun &

Wei, 2007; Braus & Elliott, 2009; Reiff & Kelly, 2010; Salzman & Bevins, 2008) Conversely, lifestyle can also alter the risk of disease. For example, smoking has shown a positive correlation particularly in CD studies. (Braus & Elliott, 2009). In essence, within the past few years, five theories has been postulated in the etiology of IBD, even though they have not all as yet been completely proven. These enumerated are; the persistence of specific infection, dysbiosis in the beneficial and dentrimental gut commensal bacteria, defective mucosal barrier function and microbial clearance and aberrant immunoregulation. (Braun & Wei, 2007; Sartor, 2006)

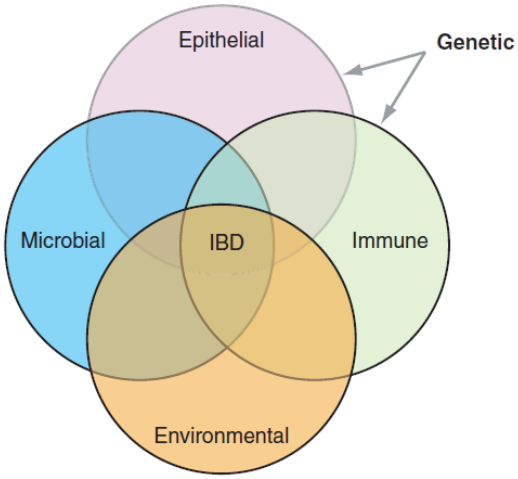

***Figure* 1.3.3**: The interacting factors in the pathogenesis of IBD. There are mounting evidence that an interplay of this factors explains the incidences of the disease in relation to control subjects (Braun & Wei, 2007)

It is believed that an interplay of the gut microorganisms, the host immune function and epithelial cells underlays an individuals susceptibility to IBD or not and by the modulation of both genetic and environmental factors (Braun & Wei, 2007; Braus & Elliott, 2009).

1.3.2.1 Host Genetic Factors

IBD (UC or CD) as a heterogenous diseases has been elucidated with a complex host genetic risk. It is envisaged that 50% of CD and 10% of UC cases are associated with genetic susceptibility (Braun & Wei, 2007). The most proxy evidence is in familial aggregation studies where, 75% patients have some first degree relatives incidence of inherited IBD. This seems to have a stronger penetrance in CD compared to UC (Xavier & Podolsky, 2007). (Braun & Wei, 2007; Braus & Elliott, 2009; Hanauer, 2006) IBD is not inherited by Mendelian trait (Hanauer, 2006), however, there are high rate concordance in monozygotic

twins. This is about 50% in CD but lower in UC and 18% than in dizygotic twins (Braus & Elliott, 2009). Lately, to highlight the genetic implication of IBD, specific genes correlating to the diseases have been identified (Braus & Elliott, 2009)

Over the last few years, genome wide screenings of single nucleotide polymorphism (SNP) and candidate genes have revealed at least 30 loci (Braus & Elliott, 2009) that contribute to the risk of IBD, most especially in CD subjects (Braun & Wei, 2007; Braus & Elliott, 2009; Hanauer, 2006; Reiff & Kelly, 2010; Sartor, 2006; Xavier & Podolsky, 2007). Similarly, some CD loci have shown to also effect the risk of UC (Braus & Elliott, 2009).

The probable most important gene so far identified on IBD-1 locus of chromosome 16 is NOD2/CARD15 (nucleotide-binding oligomerisation domain 2/caspase activation and recruitment domain 15). (Braun & Wei, 2007; Braus & Elliott, 2009) The product of this gene is a protein which encodes for an intracellular peptidoglycan receptor that activates CARD and nuclear factor (NF)-κB-dependent pathway of the innate immune cellular activation. This is important in host-bacteria recognition and clearance of bacteria, involving both proinflammatory and protective molecules. (Braun & Wei, 2007; Hanauer, 2006; Salzman & Bevins, 2008) As at now, three mutations, causing amino acid substitution has been discovered. (Sartor, 2006) Although, the biological effects of NOD2/CARD15 mutations are still unclear, its receptors localised in monocytes and Paneth cells may be responsible for the production and secretion of defensin, an antimicrobial peptide. As such can be important in the host barrier defence as shown in knock out mouse experiments and the eventually impact on the gut bacterial load. (Salzman & Bevins, 2008)

There seem to be a number of associated genes involved in the pathogenesis of IBD (Salzman & Bevins, 2008), however, defective NOD2/CARD15 genes have reported association to 17-27% of CD cases (Hanauer, 2006).

1.3.2.2 Immunoregulation

The immune reactivity of the gut mucosa to foreign agents implicated by genetic suscepitibility, is both local and systemic and humoral and cellular. (Xavier & Podolsky, 2007) Dysregulation in maintaining intestinal homeostasis is a compelling evidence in the pathogenesis of IBD from human and animal studies (Round & Mazmanian, 2009; Xavier & Podolsky, 2007). Similarly, the protective resistance of the mucosal wall is weakened in

susceptible IBD individuals leading to elevated circulating macromolecular antigens (Karlinger, et al., 2000).

As a theory, it is stipulated that normal gut flora (commensals), act in such a mechanism as to blocks proinflammatory genes that activates the NF-κB-dependent pathway and thus hinder the expression of the inflammatory adaptive immune response (Hanauer, 2006). Hence, in IBD this mechnaism has been lost through infection of pathogenic bacteria (Karlinger, et al., 2000) or alteration of "normal" bacterial function which eventually leads to chronic and destructive immune responses (Hanauer, 2006). Although, the excessive activation of adaptive immune response is directly implicated in tissue damages (inflammation) in IBD patients. The innate immunity is the driving proxy of the residents and recruited mucosal B and T cell population. (Xavier & Podolsky, 2007) Conversely, risk of CD has been elucidated by the identification of mutations in the autophagy-related gene ATG16L1. This leads to impaired innate immune defences by as yet unknown mechanisms. (Braus & Elliott, 2009)

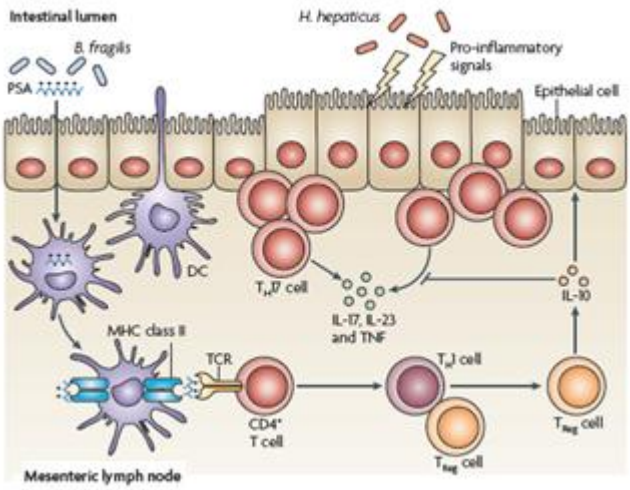

***Figure* 1.3.4**: Mechanism by which T_{reg} cells exerts gut immunological homeostasis. Proinflammatory cytokines are thought to be downregulated by inhibiting T regulatory cells by mechanisms not completely elucidated. (Round & Mazmanian, 2009)

Within IBD adaptive immune system, immunomodulation comprises of the immunglobin A and G produced in the mucosal B cells and the complex mixture of T cells lymphocytes; T helper 1 (T_H1), T helper 2 (T_H2) or T helper 17 (T_H17) phenotypes (Xavier & Podolsky, 2007) which produces proinflammatory cytokines; interleukin (IL), tumour necrosis factor (TNF) and interferon-γ (IFNγ) (Round & Mazmanian, 2009). A consensus from various

studies has indicated that CD is characterised by T_H1 mediated inflammation responses, whereas UC exhibits T_H2-like cytokine profiles (Braus & Elliott, 2009; Xavier & Podolsky, 2007) From animal and human studies, evidence exists that suggests that T_{reg} cells are implicated in regulating intestinal immunological balance in experimental UC. As shown with mechanism of action illustrated in *Figure* 1.3.4, *Bacteroides fragilis*, is demonstrated to illicit mucosal protective action on proinflammatory signals from *H. hepatcus* on epithelial cells through the secretion of polysaccaride A (PSA) and presentation on MHC class II of CD4 T cells. (Round & Mazmanian, 2009; Xavier & Podolsky, 2007).

In general, there seem to exists more convincing data in the immunological pathogenesis of CD while the role of antigens and antibodies in UC is yet to be determined (Karlinger, et al., 2000). Furthermore, an over stimulation of the immune responses in susceptible individuals can also be attributed to IBD pathogenesis (Bouma & Strober, 2003).

1.3.2.3 Environmental and Lifestyle Risk Factors

As mentioned earlier, the higher prevalence of IBD in developed and industralised regions such as the Scandinavia, and the development of IBD by migrants from rare incidence regions gives weight to the theory of environmental influences in IBD development. Thus, IBD has be perceived with lower incidence in regions of poorer sanitation and hygiene. (Hanauer, 2006) Often hinging on ones socio-economic status, the contributing lifestyle environmental factors includes, diet, smoking, exposure to sunlight, sanitation, chemicals (industrial/drugs), and occupation. (Hanauer, 2006; Karlinger, et al., 2000)

A link between dietary factors and diseases is thought to be implicated with the number of diseases appear in the GIT (Karlinger, et al., 2000). However, studies linking diet to IBD incidence are still inconclusive. (Hanauer, 2006) In an instance, a 3 to 4-fold greater risk has been projected in frequent consumers of fast-food, a suggestion that the westernised dietary intake of high fatty acids have been associated with the risk of IBD. Other food substances, alcohol and size of of particles in the gut content has been reported in persistent CD. (Karlinger, et al., 2000)

The observation relating smoking and IBD are complex but a unique pathophysiological factor (Hanauer, 2006; Karlinger, et al., 2000) which may overide genetical factors (Hanauer, 2006). In UC from diverse geographical regions, smoking is shown to be rather

protective in a dose dependent manner (Hanauer, 2006) but has a negative influence (Karlinger, et al., 2000) on CD patients. Conversely, ex-smokers are even 1.7 times at risk of IBD compared to those who has never smoked and twice likely to require colon surgical procedure (Hanauer, 2006).

1.4 Gut bacteria and IBD

The diverse human gut microbiota contains about 10^{14} microbes of greater than 500 species. In the course of time, many different bacteria has been hypothesied as causative agents of IBD in addition to other etilogical agents (few enumerated earlier). This may be due to the resemblance of symptoms of IBD with certain bacteria infections, such as in the early symptoms of Yersinia infection that is similar to symptoms of CD (Karlinger, et al., 2000). This theory is been applied to experimental evaluation, especially with the advent of the robust metagenomic approach to enumerating previously un-cultivable gut bacteria. However, results has not been consistent in pinpointing a specific bacteria agent in IBD etiology. Nonetheless, in susceptible individuals, pathogenic or normally enteric bacteria may induce and perpetuate chronic gut mucosal inflammation. (Sartor, 1997) UC is not observed in mice studies within a germ free environments (Bouma & Strober, 2003), hence, the theory implicating the human gut microflora in IBD may be valid (Hanauer, 2006).

Consistent with most metagenomic studies, in health, the human gut microbiota is dominated by the members of the Bacteriodetes and Firmicutes phyla (Ley, et al., 2006; Mai & Draganov, 2009) From various IBD studies, etilogically implicated bacteria include, *Faecalibacterium prausnitzii* exerting a positive impact, described as antiinflammatory, while *Escherihia coli* and *Mycobacterium avium paratuberculosis* (MAP) have a negative impact as potential IBD infectious agents (Mondot et al., 2011; Salzman & Bevins, 2008; Sokol, Lay, et al., 2008; Sokol et al., 2008; E F Stange, et al., 2006; E. F. Stange, et al., 2008). These are mainly established from differences in gut bacterial composition in analysed feacal samples between observed IBD cases and healthy subjects (Mai & Draganov, 2009).

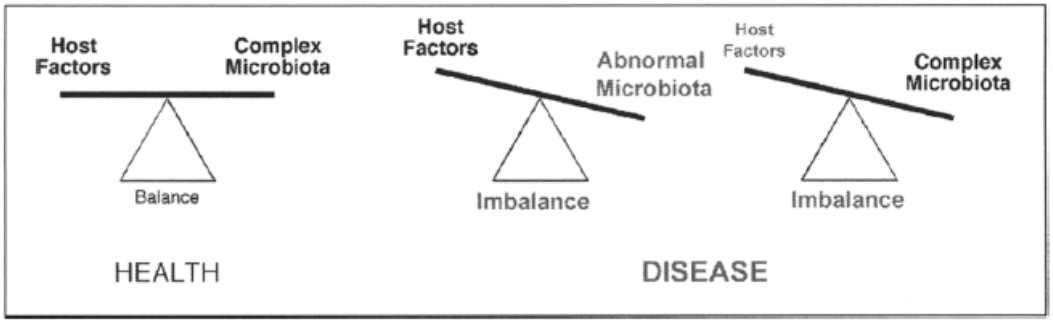

***Figure* 1.4.2**: Model showing the schematic interaction between the host factors and the microbiota in disease (imbalance) and health (balanced). (Salzman & Bevins, 2008)

Thus, the interactive relationship, symbiosis or pathogenic, between the host and the gut bacteria are shaped by selective pressures within the host (genetic) and competitive modulation of resident microbial community, of which net effect may have implications on the health of the host. (Ley, et al., 2006; Ventura, et al., 2009) Similarly, an imbalance in the composition of the commensals or beneficial (symbiotic) bacteria and pathogenic bacteria creates an abnormal host microbiota, which may lead to IBD diseased state in risk bearing individuals as illustrated in *Figure* 1.4.2. (Blaser & Falkow, 2009; Round & Mazmanian, 2009; Salzman & Bevins, 2008; Turnbaugh et al., 2007)

The gut dysbiosis observed could be an increase or a reduction in diversity of a group or specific bacteria. However, a drawback in attributing microbiota composition to disease is that is not a concensus regarding the normal range of microbiota diversity on a population level to makes this conclusions (Mai & Draganov, 2009).

1.5 Describing Microbial Diversity

An understanding into the role of the human gut microbiota is impeded with our limited knowledge of the milieu, individuality and which are often too conceptualised. The functional elucidation of the numerous and incredibly complex human gut microbiota is quiet labourous, technically intensive and expensive. The inaccessibility of the gut means the faster, easily reproducible *in vitro* approaches forms the most palatable, comparable to *in vivo* strategies. Despite, it is difficult mimicking *in vivo* colonic conditions exactly. Culturing approaches has yielded limited success given that approximately 80% of the species residing in the human gut are un-culturable obligate anaerobes. (Brugère, Mihajlovski, Missaoui, &

Peyret, 2009; Kovatcheva-Datchary, Zoetendal, Venema, de Vos, & Smidt, 2009; Peterson, Frank, Pace, & Gordon, 2008; Reiff & Kelly, 2010)

Conversely, culture-independent studies based on nucleic acid identification are breaking grounds in realising the impact of the gut microbiota on the host health. (Reiff & Kelly, 2010) The current metagenomics approach has improved the technicality of isolating diverse bacteria in specific communities and thus present a phylogenetic picture of species richness. (Reiff & Kelly, 2010; Tuohy et al., 2009) However, as the traditional culture methods improve for previous uncultured bacteria (Peterson, et al., 2008), a blend of the both strategies may completely establish the physiology, functionality and abundance of the metabolic active human gut microbiota species and its impact on the host responses (Brugère, et al., 2009; Reiff & Kelly, 2010).

In details, the gut microbial community can be enumerated by applying some of the high-throughput nucleic acid-based strategies in PCR amplification and PCR-based analysis, cloning, and sequencing among others. (Brugère, et al., 2009; Leser et al., 2002; Nocker, Burr, & Camper, 2007)

1.5.1 Molecular markers and Amplification

Bacteria classification has basically been by the nucleic-acid identification using the small subunit (SSU) of ribosomes, particularly, the 16S rRNA gene (Brugère, et al., 2009), often termed as universal molecular chronometer (Woese, 1987), present in all cellular organisms. These genes are highly conserved and offers a quick and orderly way of identifying gut microbe phylogenetic diversity (phylotype) and variations in the sequences (Achtman & Wagner, 2008), probably down to the species level (Brugère, et al., 2009) and may reliably indicate relatedness in the absence of lateral gene transfer (Peterson, et al., 2008). Thus, microbial species maybe defined despite no acceptable consensus (Achtman & Wagner, 2008; Peterson, et al., 2008) as strains with greater than between 97% to 99% 16S rRNA identity or more than 70% DNA-DNA hybridization (Achtman & Wagner, 2008; Brugère, et al., 2009; Peterson, et al., 2008) are classed as species. Furthermore, the conserved 16S rRNA regions are used in microbial studies as they are relatively small, approximately 1.5 kb, and can easily be target amplified by polymerase chain reaction (PCR) with identified universal oligonucleotide primers within a sample pool of different microbes, even at minute

quantities (Peterson, et al., 2008) and also useful in oligonucleotide probe targets (Brugère, et al., 2009).

This molecular tool has greatly advanced attempts to elucidate the proposed implication of microbial factor on IBD (Reiff & Kelly, 2010).

1.5.2 PCR-Based Analysis

PCR-based fingerprinting as a molecular in tool in microbial community identification has proved useful, providing an overview of the resident gut bacteria and monitor dynamics in richness (Tuohy, et al., 2009). However, apart from DNA extraction protocol bias that can be carried over (McOrist, Jackson, & Bird, 2002; Trosvik et al., 2007), PCR efficiency bias of mixed template differential amplification (Trosvik, et al., 2007) and at other downstream stages (Tuohy, et al., 2009), which may leads to relative under or over representation of gut microbiota taxon. If this bias are overcomed, its been an invaluable tool for microbial community profiling. (Brugère, et al., 2009)

Based on similar techniques, one of most commonly used PCR-based analysis protocol are the denaturing-gradient gel electrophoresis (DGGE) and temperature-gradient gel electrophoresis (TGGE) and its derived variants. This method reveals profiles of differential separation, analysing the 16S rDNA amplicons by denaturing rate of dsDNA, DGGE or temperature-dependent migration rate, TGGE. It is rapid and cost effective and may expose temporal flunctuation within the gut microbiota (Alper et al, 2005) and impact of environmental consequences when optimized (Tuohy, et al., 2009). Furthermore, terminal-restriction fragment-length polymorphism (T-RFLP) is another molecular technique for profiling microbial communities and has been applied in digestive environment to quantify gut microbial diversity based on terminal DNA fragments variants of known enzymes restriction sites of 16S rRNA gene using fluorescent labelled PCR primers (Brugère, et al., 2009). Similarly, the use of the automated ribosomal intergenic spacer analysis (ARISA), can discriminate closely related species using the 16S rDNA analysis. (Achtman & Wagner, 2008; Nocker, et al., 2007) This uses the high heterogenic length and nucleotide sequences of the intergenic transcribed spacer (ITS) region between the 16S and 23S ribosomal genes. In this process of microbial community profiling, the fluorescently tagged PCR product is detected on a laser sensitive automated sequencing system. (Nocker, et al., 2007)

1.5.3 Cloning and Sequencing

The cloning strategy is a step in the identification of gut bacterial species by sequencing. Clone libraries are generated from PCR products of bacterial 16S rRNA pool, sticky ends such as 3'A overhang, generated by *Taq* polymerase is ligated to an efficient vector as inserts (Nocker, et al., 2007). In this way, an inventory, proximate of the microbial diversity in a community can be achieved from the sequences clone library (Leser, et al., 2002; Nocker, et al., 2007). However, the cloning process can be tedious and time consuming but can be cost effective with advances in high-throughput sequencing.

In the last two decades sequencing of conserved genes has been used in phylogenetic tree construction and comparative studies of the gut microbiota and as served as an alternative to some PCR-based analysis which can be quite laborous (Nocker, et al., 2007). Typically, the classical sanger sequencing based on the fluorescently labelled DNA chain-terminating dideoxynucleotide triphosphates (ddNTPs) on the single base resolution of electrophoretically separated deoxyribonucleotide triphosphate fragments, has been efficiently used to obtain DNA sequences. Unfortunately, its expensive and limiting in throughput. (Brugère, et al., 2009; Peterson, et al., 2008)

The latest and future sequencing technologies are poised to circumvent this limitations and as now applied in metagenomic analysis. This includes, pyrosequencing and the improvement in 454 pyrosequencing, that relies on sequencing by synthesis, with the later using immobilised and amplified templates on aqueous oil emulsified beads which can generate up to 400 bases as opposed to 100 bases by the pioneering pyrosequencing. By this, sequences are detected using a chemiluminescence enzyme tracking of DNA polymerase synthesis activity in the release of pyrophosphate on nucleotide incorporation (Brugère, et al., 2009) Furthermore and among others, analysis involving sequence by hybridisation for genotyping that uses oligonucleotide arrays is the future of high throughput DNA sequencing as it excludes most repetitive sequences and can by-pass the cloning activity of bacterial 16S rDNA which can be efficient for DNA fragments that cannot be propagated in bacterial vectors. Although, sequencing by synthesis can also be applied on oligonucleotide arrays. (Brugère, et al., 2009)

Capillary gel electropheresis has suitably been used in the detection and separation of as far a single nucleotide differences in DNA and genomic sequencing. This tool is attractive

compared to convectional gel electrophoresis as more than 600 nucleotides can be resolved in one automated run. This DNA analysis by sequencing technology is faster and efficiently resolves DNA fragments, utilising high voltages of between 10-30 kV in fused 10-100 μm diameter silica micro-capillaries and fixed to ultra-sensitive fluorescence or UV absorbance detectors. (Cohen et al., 1988; Swerdlow & Gesteland, 1990)

1.5.4 Microarray

The use of DNA microarray has been an effectively powerful experimental tool used to simultaneously monitor several thousands of genes, measuring abundance and diversity in a community and often optimised for gut microbiota in IBD diagnostics platforms (Brugère, et al., 2009; Kovatcheva-Datchary, et al., 2009). The microarray platform is designed with taxonomically differentiated specific DNA sequences labelled probes that will hybridise to target DNA fragments on especially planar glass surfaces. This can be sensitive to single nucleotide polymorphisms (SNP) on target species but cross-hybridisation can be an issue. However, it has been used to analyse relationships between healthy and diseased gut microbiota (Kovatcheva-Datchary, et al., 2009). Flow cytometry can also be used in the analysis of microbial community similar to the microarray principle above, its limiting on the number of phylotypes that can be sorted at the same time. Furthermore, microarray technology has been advanced in the implication of microbial metabolic activity in the gut disorders, using radioistope labelled DNA or RNA probes to provide phylogenetic information. (Brugère, et al., 2009; Kovatcheva-Datchary, et al., 2009)

Although, microarray represents an excellent tool to analyse and expose potential diagnostic biomarkers, but for a larger clinical scale analysis, combining the FlowMetrix system from Luminex is thought to be simpler, flexible, cost effective and high-throughput for the validation of biomarkers. The Luminex can thus, screen upto a hundred different transcripts in many samples by a multiplex of ligation mediated amplification. (Peck et al., 2006; Sherry A, 2006) This xMAP® technology, is fully automated, lists its application within a variety such as measuring cytokines in serum, transcription factors in cell lysate, or detecting SNP (single nucleotide polymorphs) and has been applied to microbial detection even at low DNA concentrations and diversity studies (Sherry A, 2006). It performs discrete bioassay on the surface of microspheres, which are colour coded beads of two flourescent dyes with 100 different intensities for analyses. Multiple lasers, report by identifying the colourings on each individual microspheres particles from reads on the compact analyser.

1.5.5 Specimen materials for mining gut microbiota

In IBD studies, at least 50% intestinal bacteria is contained in the faecal matter (stool) (Braun & Wei, 2007) and biopsis have provided the rich wealth of gut microbes for it characterisation. (Brugère, et al., 2009; Lepage et al., 2005) Nonetheless, bacterial composition in the various intestinal sections are not represented by faecal bacterial composition and neither does it represent mucosal bacterial composition. (Salzman & Bevins, 2008)

1.5.6 Data and Statistical Analysis, Bioinformatic tools

In the rapidly evolving knowledge into microbial ecology and microflora implication to host dysregulation, the large accumulation of molecular data in predicting diversity, abundance and novel microbial function requires a theory of predictive power that is of practical value. The value of this theory to the microbiologist will aid the interpretation, classification and prediction of richness as observed through experimental data. (Prosser et al., 2007) However, most important theories may have limited application to microorganism and especially the gut microflora (Achtman & Wagner, 2008; Prosser, et al., 2007) and may not be suitable for high-throughput application (Trosvik, et al., 2007). Exploring the usefulness of metagenomic approach to unraveling microbial communities, computational tools have thus been developed to streamline and distil interpretation of the numerous microbial data post sequencing (Kunin, Copeland, Lapidus, Mavromatis, & Hugenholtz, 2008; Peterson, et al., 2008). With most of the culture-independent sequences utilising the SSU rRNA (I6S), the Ribosomal Database Project (RDP) analysis has been used for taxonomic visualisation and to link phylogeny. The RDP alignment strategy includes as at date, greater than a million quality-controlled bacterial and archaeal small-subunit 16S rRNA sequences. (Cole et al., 2009) However, other databases such as greengenes also provide current and comprehensive 16S rRNA gene sequence alignment for phylogeny browsing, blasting and probing (Brugère, et al., 2009) and can further assist the researcher in interpreting microarray results and annotating novel sequences. Recent advances in microbial bioinformatic tools has created the MEGAN, a BLAST that profiles community composition based on sequence similarities with the highest hit, assuming the nearest phylogentic match by the lowest common ancestor algorithm (Kunin, et al., 2008).

Statistical analysis of metagenomic data can be useful to extrapolate frequency dynamics in the structure of microbial communities as in higher organisms. This way, bias mainly from probe and clone library data sets in DNA sequence alignment and taxonomic inference maybe by-passed. (Knut Rudi et al., 2007; K. Rudi, Zimonja, Trosvik, & Næs, 2007) In the face of the large amount of data from microbial community studies, multivariate curve resolution (MCR) is a mathematical tool with the goal to expose the individual elements effects in a mixture of components. (Garrido, Rius, & Larrechi, 2008) In general, using multivariate statistics is a way of finding structure in a data with lots of variables in a condensed format. One of such application is the principal components analysis (PCA), which gives a small number of linearised combination of the variables in that data set that highlights the existing variations in the whole data frame.

1.6 Potential of gut bacteria in IBD Diagnostics and Management

Currently, there seem to be no single pathogenic marker for IBD and as such approach to diagnosis involves a combination of clinical presentations, biochemical features and pathological examinations (Baumgart & Sandborn, 2007). The two forms of IBD, UC and CD can thus be confirmed (Baumgart & Sandborn, 2007) or differentiately diagnosed by the evaluation of a combination of this features manifested, however, this can be sometimes not so distinctive (Hendrickson, Gokhale, & Cho, 2002).

Presence of blood and mucus in stool are seen in intestinal clinical presentation, which seems to be diagnosed early in UC compared to CD because of gross blood in stool (Hendrickson, et al., 2002). The extent and severity of blood and frequency of stool can classify disease. Conversely, the presence of diarrhea with blood and mucus exclude the similar but different inflammatory bowel syndrome (IBS) in IBD diagnosis (Baumgart & Sandborn, 2007). Abdorminal pain are experienced and can intensify in UC during bowel emptying. Extent and localisation and progression of the pain can differiate CD from UC with pains in left lower quadrant extending to the entire colon in UC patients and right lower abdominal quadrant in CD suffers. (Hendrickson, et al., 2002) Partial or complete colonic obstruction, stenosis with anal fissures and fistulae is a more common manifestaiton in CD patients. Also, extraintestinal markers observed in IBD sufferers include fever, weight loss especial in CD, delayed growth and sexual maturation in children, ulcer, osteopenia and renal implications. (Baumgart & Sandborn, 2007; Nocker, et al., 2007)

Biochemical tests such as stool examination is very paramount to exluding enteric pathogens exhibiting similar IBD symptoms such as *Salmonella*, *Shigella*, *Yersinia*, *E. coli*, *Mycobacterium tuberculosis* among others but will persist despite applying normal treatments for such infections. This test can also be performed in events of relapse (E. F. Stange, et al., 2008). Serological test investigates full blood count, microlytic anaemia, leukocytosis and thrombocytosis and results may give an indication of IBD. Low iron levels and high liver enzymes are also observable as biochemical inflammatory marker, such as c-reactive proteins (CRP) which are less in UC but seen in elevated levels in 90% of CD patients. Perinuclear anti-neutrophilic cytoplasmic antibody (P-ANCA) and anti-*Saccharomyces cerevisiae* antibodies (ASCA) test are useful to support IBD diagnosis and ASCA was seen to be highly specific when it presented in 89 to 97% of children with CD. (Baumgart & Sandborn, 2007; Hendrickson, et al., 2002) Genetic marker test may also be performed, although not conclusive in UC, NOD2 mutations of particular phenotypes have been associated with CD susceptibility and screening. Elucidation of other implicated genes are in the horizon (E F Stange, et al., 2006; E. F. Stange, et al., 2008).

The suggestion of IBD manifestation can be further established by endoscopic and radiologic evaluation. Endoscopy and biopsis can be particularly useful in differentiating between UC and CD, while equally useful in determining the extend of colonic involvement in CD sufferers, utilising barium enemas. (Hendrickson, et al., 2002) These diagnostic features show at least 50% sensitivity and specificity and are considered due to its moderate reproducibility (E F Stange, et al., 2006; E. F. Stange, et al., 2008).

With advancing knowledge in the etiology and prognosis of IBD, management has been based on a concensus of suggestive therapies which are not essentially curative but to maintain remission and flare ups (Baumgart & Sandborn, 2007). IBD therapies has mainly been through drugs adminstration to induce or prolong remission, nutritional intervention to control the disease and finally surgery for either or both of the types of disease, UC and CD.

The choice of drug therapy is dependent on the severity of the symptoms and the site and degree of disease involvement. Conversely, adverse side effect may be inevitable with some of the medications. In mild UC cases, sulfasalazine alone or combination with topical medications is used but intolerable side effects in some patients are overcomed using the newer 5-aminosalicylic acid medications. In CD cases, corticosteroids are used in the short term due to undesirable side effects for therapy to induce remission and disease activity.

However, in mild and moderate CD diseases, sulfasalazine is used to maintain remission induced by corticosteroids. Also in both diseases, immunosuppressive therapy are often administered using steroid free azathioprine or 6-mercaptopurine and for faster effects in achieving or prolonging remission. Similarly, antibiotics such as metronidazole and ciprofloxacin are administered to CD patients at least in the short term. (Baumgart & Sandborn, 2007; Hendrickson, et al., 2002)

Biologically, due to implication of TNF-α in CD pathogenesis, infliximab, an infused monoclonal anti-TNF-α have been successful in inducing and maintaining remission in 48% of cases. Furthermore, thalidomide, despite its side effects are effective inhibitors of TNF-α. (Hendrickson, et al., 2002)

Finally, surgery is used to treat patients with very severe cases of perforation, rectal bleeding, toxic megacolon and irresponsive to medical managements and in complications. (Hendrickson, et al., 2002)

1.7 Project Rationale and Aim

Idiopathic IBD has posed a challenge for complete management. Human gut bacteria on the evidence of recent studies, have been implicated in the pathogenesis of the diseases. Thus, there is need for early diagnostics approaches, an important step in achieving needed therapy. However, this target has fallen short largely due to complexities of the gut microbiota and the applicable tools for the elucidation of highly un-culturable intestinal microbes. To completely understand the complex changes in the gut microbiota composition that may predispose inflammation or even promote health, theres need to pursue assay techniques that can unravel any link between the metabolic activities of the bacteria in the gastrointestinal tract and diversity.

In this project, the aim was to determine correlations between the composition of the gut microbiota and IBD in children (<18 years olds) and evaluate potentials for diagnostics. The approach involved molecular techniques including,

- Direct sequencing of purified and quantified pool of bacterial 16S rRNA gene from children stool samples
- Screening which incorporates Genetic Analysis (GA) AS novel gut microbiota probe tool
- Enumeration of bacterial diversity and phylogeny from generated clone library.

This project in comparison to previous studies found in literature, seem to be one of its kind trying to extrapolate dynamics in gut microbiota residents in IBD subjects with greater than 50 patients stool samples, early diagnosed and untreated cases.

1.7.1 Outline of the Project

This project is a 60-credit work, constituting the MSc Thesis of the Hedmark University College Masters programme in Commercial and Applied Biotechnology, 2011. The project was involved a collaboration between Genetic Analysis AS (GA) and Akershus University Hospital (Ahus). With a large bio-bank of faecal material at Ahus in the IBESEN II study, and applying the GA novel gut probes screening technology this project was achieved.

A flow scheme of how the outline of this project was achieved is shown below:

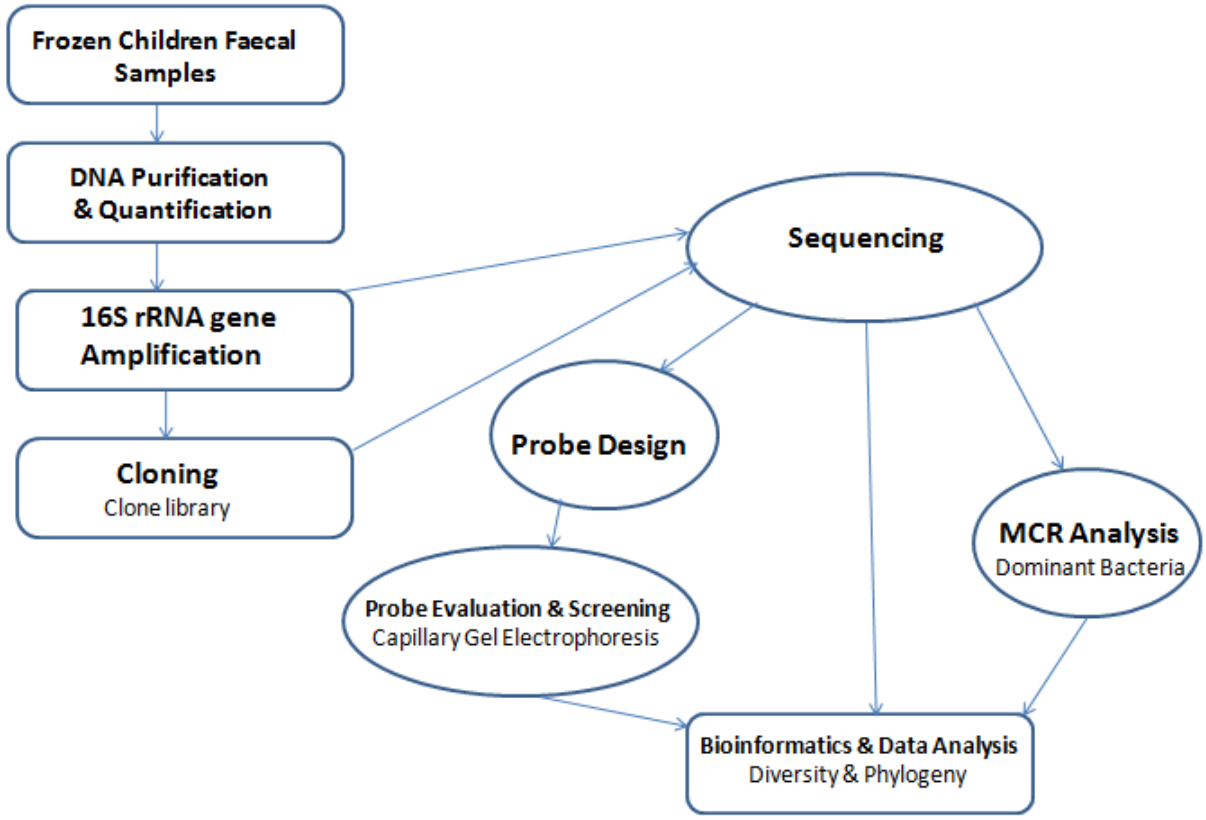

Figure 1.7.1: Schematic outline of the project work. In achieving the aim of this project work, analysis involved purification, quantification and amplification of DNA from 75 children stool specimens. Identification of faecal bacterial load was accessed by sequencing, multivariate curve resolution analysis, cloning and sequence bioinformatics analysis. Probe evaluation and screening of the amplified 16S rRNA gene and data subjected statistical analysis was essential to enumerate potential for IBD diagnostic.

2. Materials and Methods

2.1 Samples Cohort (IBESEN II)

A total of 75 children samples (<18 years old) stored at -80°C, were provided from diagnosed patients stool specimens. Samples were deposited from diagnosed, early inflammatory bowel disease (IBD) patients at Akershus University Hospital (Ahus), Oslo Norway. These were part of the Norwegian IBSEN II study. From these collections, 27 were diagnosed Crohn's Disease (CD), 16 diagnosed Ulcerative Colitis (UC) and 30 samples were from diagnosed non-IBD subjects (control) and with 2 "unconfirmed" prognosis (IBDU). Full list included in *Table* 2.1.1A.

IBD diagnosis criteria for patient specimen included in the IBSEN II cohort were, abdominal symptoms including diarrhea and/or blood in stool for more than 10 days, endoscopic or radiological examinations for signs of inflammation and histological signs of chronic inflammation.

Subjects with pathogenic gut bacteria infection (except *Mycobacterium avium*), parasites, cysts and eggs were excluded from this cohort. Similarly, comorbid patients with cancer, haematological or hepatological disorders, significant cardiovascular, neurological and respiratory conditions were not included in these study. In addition, other chronic inflammation were exempted from this study in both disease and control subjects.

Finally, IBD diagnosed patients were classified as CD, UC or IBDU based on satisfying the set criteria (Table 2.1.2A in Appendix).

2.2 DNA Purification and Quantification

Mechanical lysis was achieved as per Ahus DNA prep protocol. Samples were thawed from frozen, -80°C and a scalpel, 180 to 220 mg of stool in each 2 mL microcentrifuge tube. These were mechanically and vigorously lysed with 1.6 mL of ASL buffer (Qiagen, Hilden, Germany) for 2 minutes at 30 Hz using magnetic beads (Qiagen, Hilden, Germany) on Qiagen® TissueLyser (Qiagen, Hilden, Germany) and further lysis at 95°C for 5 minutes in a heating block. On slight cooling, tubes containing lysed samples were centrigude at 17 G for 15 seconds and again at 17 G maximum speed for 1 minute. To the 1.4 mL of pipetted

supernatant in a fresh 2 mL tubes, and inhibitorEX-tablets (Qiagen, Hilden, Germany) were added with 1 minute bench incubation and 3 minutes maximum speed centrifugation at 17 G. Final centrifugation at 17 G for 3 minutes was repeated with at least 600 µL of supernatant in a new 2 mL tube.

Similarly, a final automated purification of the DNA from the 600 µL supernatant lysate were achieved using the QIAGEN kit (Qiagen, Hilden, Germany) on the QIAcube purifier(Qiagen, Hilden, Germany). These protocol were as described in the QIAamp® DNA Stool Handbook (the relevant protocol sheet on www.qiagen.com/MyQIAcube). The QIAcube purifier were preloaded with reagents; proteinase K, AL buffer, ethanol, AW1 and AW2 buffers and AE elution buffer. The run selection on the QiaCube were as programmed at Ahus for stool DNA extraction: (ON) > DNA > QIAmp DNA stool > Human stool > Human DNA analysis > Start.

Determination of concentration, yield and purity of extracted bacteral DNA pool from stool samples were achieved using the Nanodrop spectrophotometer, calculating the ratios at $Ab_{260/280}$. Similarly, DNA length was visualised through 1.5% (w/v) agarose gel electrophoresis, stained with ethidium bromide at 100 V for 30 minutes using 100 bp or 1 kb ladder and 2 µL of purified DNA and 1 µL of loading buffer per well and ran on 1X TAE buffer. This conditions were used in subsequent gel electrophoresis analysis.

The purified DNA eluates were all normalised to a final working concentration of 1 ng/µL for all subsequent assays and analysis. This was determined by a dilution series of; 1, 1/10; and 1/100 and PCR gel image strength empirically selected from stock DNA samples extracted; two of each strong, weak and moderate visualised gel bands by subsequent PCR (GeneAmp PCR Systems 9700, PE Applied Biosystemss, Norwalk, USA or AB Applied Biosystems 2720 Thermal Cycler, Singapore) amplification with primers and parameters as outlined below and analysis of the product on 1.5% (w/v) gel Red® stained agarose gel.

2.3 Polymarese Chain Reaction (PCR) amplification of 16S rRNA

Using a combination of forward; 5'-TCCTACGGGAGGCAGCAG-3' (MangaF-1) and reverse; 5'-CGGTTACCTTGTTACGACTT-3' (16S-UR) primers generally targeting the conserved 16S rRNA DNA regions of multiple microorganisms, approximately 1200 bp, was used in the amplification. The reaction mix on ice contained 1.25 U HotFirePol (Solis

Biodyne, Tartu, Estonia), 1XB2 buffer(Solis Biodyne, Tartu, Estonia), 2.5mM $MgCl_2$ (Solis Biodyne, Tartu, Estonia); 200 µM dNTP (Thermo Fisher Scientific, Surrey,, UK); 0.2 µM each of forward [Mangala F-1] and reverse [16S-UR] primers to 1 µL of 5 to 100 ng/µL (w/v) DNA template to a final volume of 25 µL. The PCR thermocycler was programmed for initial denaturation at 95°C for 15 minutes, and with 30 cycles of denaturation at 95°C for 30 seconds, annealing for 30 seconds at 55°C, elongation for 1 minute 20 seconds at 72°C and at the end a final elongation for 7 minutes at 72°C.

For the PCR analysis of cloning transformants however, the forward and reverse, primers; 5'-CGCCAGGGTTTTCCCAGTCACGACG-3'(HU) and 5'-GCTTCCGGCTCGTATGTTGTGTGG-3'(HR) respectively were instead used to target specific regions of approximately 1400 bp of the 16S rRNA gene with the flanking regions on either side.

Furthermore, PCR product were visualised on 1.5% agarose gels, prior to any further analysis and probably for confirmation of amplification.

2.4 Sequencing

In an attempt at determining the identity of the bacterial pool of the 16S rRNA gene by direct sequencing, the PCR product were firstly cleaned up of excess nucleotides. By treating with 3U Exonuclease I (ExoI) and 8U SAP, Shrimp Alkaline Phosphatase (usb Corp, OH, USA) at 37°C for 2 hours and inactivated at 80°C for 15 minutes.

The ExoSAP treated PCR product were diluted 10X, 1 µL placed in each well and included with 0.32 µM each of 5'-[C X30]CGTATTACCGCGGCTGCTGGCAC-3' (U515FC30) primers, 1X BigDye buffer and 1 µL BigDye v1.1, incorporation reaction to a 10 µL total volume. The PCR thermocyler was programmed at 25 cycles of denaturation at 96°C for 15 seconds and annealing and elongation at 60°C for 4 minutes. Similar approach in sequencing the picked clones with insert used using the 16S rRNA specific primers, 0.32 µM of forward; 5'-TCCTACGGGAGGCAGCAG-3' (MangaF-1) and reverse; 5'-CGGTTACCTTGTTACGACTT-3' (16S-UR) primers instead.

BigDye® Terminator v1.1 cycle sequencing Kit (AB Applied Biosystems, CA, USA) was added to above completed ExoSAP reaction, volume of 10 µL. According to manufacturer's instruction this addition was a 55 µL mastermix of Xterminator which included 10 µL

Xterminator and 45 µL SAM. These were vortexed (IKA® MS 3 digital, USA) with shaking at room temperature for 30 minutes at 1500 rpm and centrifuged (IEC Centra CL3, USA) at 1078 G for 2 minutes. Sequences analysis was on the ABI Genetic Analyzer 3130xl sequencer with 36 cm capillary array containing optimized polymer 7 (POP-7, Applied Biosystems). Injection time was set at 6 seconds at 90°C. The sequences generated were based called by the Sequence Scanner Software v1.0 (Applied Biosystems) and used for further bioinformatic analysis.

2.5 Cloning

Using the MCR resolved sequence components, DNA pools were empirically selected for cloning from each component, 15 samples in total. Corresponding to 3 for each component and within each classification of UC, CD and Control.

2.5.1 Transformation

The experimental outline used in generating the clone libraries of pure bacteria from direct sequencing was as stipulated in the TOPO®TA Cloning Kit, Invitrogen (http://tools.invitrogen.com/content/sfs/manuals/topota_man.pdf). Transformations was achieved into the plasmid vector, One Shot® Competent Cells; Chemical or Electrocompetent (*E.coli*). Fresh PCR products were generated for each cloning reaction as in PCR amplification of 16S rRNA stipulated earlier. TOPO®TA Cloning reaction, 6 µL comprising of 4 µL fresh PCR product; 1 µL Salt Solution (Chemical Cloning) or Dilute Salt Solution (Electroporation); 1 µL TOPO® vector were mixed gently and incubated for 30 minutes at room temperature and placed on ice. For the transformation of the competent cells, 2 µL TOPO®TA Cloning reaction were inoculated into each frozen (-80°C) and ice-thawed vial of competent cells with gently mixing. Incubation was at room temperature for 30 minutes and 42°C water bath heat-shock for 30 seconds and speedly placed on ice chemically competent cells or pulsed twice in ice cooled electro-cuvettes on Electroporator 2510 (Eppendorf, Nether-HLnz GmbH, USA) at 1250 V for electrocompetent cells. Transformed cells are place in 15 mL snap-cap "Falcon" tubes with 250 µL room temperature SOC medium and incubated (innova™4080 incubator shaker, New Brunswick Scientific Co. Inc, New Jersey, USA) at 37°C for 1 hour at 200 rpm. The resulting broth, 20/40 µL was spread out on a 30 minutes pre-warmed 50 µg/mL kanamycin selective LB agar plate, pH 7.0; 1.0% Tryptone; 0.5% Yeast Extract and 1.0% Sodium Chloride (NaCl),

autoclaved at liquid cycle for 25 minutes at 121°C and cooled prior to pouring on 10 cm plates (30 – 50 mL) inverted and stored at + 4°C in the dark. The surface of the LB agar plate were tinted with 40 mg/mL X-gal in dimethylformamide (DMF) prior to pre-warming and plating of the transformed cells.

The kanamycin selective LB agar plated transformed cells were incubated (Termaks TS4115, Bergen Norway) at 37°C for at least overnight and colonies, white or light blue (as positive transformants) were picked with looped plastic sticks and resuspended individually in 50 μL 1X TE buffer and lysed at 99°C for 5 minutes. The recovered supernatant, from 4000 rpm 3 minutes centrifugation of lysate was stored and subjected to further analysis.

Conversely, a transformation control reaction was included in the cloning strategy experiments in the first instance to establish efficiency. The was same as the strategy for cloning samples above, however, the pUC19 plasmid was used instead of TOPO® vector and the selective LB agar plate contained 100 μg/mL ampicillin.

The analysis of the positive transformants was achieved by PCR and gel electrophoresis. The pure sequences were elucidated by sequencing from based called spectra.

2.6 Probe Design

Sequences that were devoid of chimeras from the clone library were trimmed to similar length within the conserved regions of the 16S rRNA gene to fit the Genetic Analysis (GA) in-house-developed PhyloMode programme used in the design and construction of the probe. Firstly, sequences were transformed into multimers (5mers), normalised and compressed into sequence related taxa clusterings plot by Principal Component Analysis (PCA).

Subsequently, the PCA plots were exported as a “pcam” file and subjected to another GA in-house TNTProbe programme that identifies probes statisfy the criteria for defined targets detection and non-targets exclusion based on combined criteria of hybridisation and labeling. All probes were designed with minimum Tm of 60°C by the nearest neighbor method for the target group, while the non-target should have a Tm of < 30°C, or the absence of a cytosine as the nucleotide adjacent to the 3’-end of the probe. All probes satisfying the criteria were identified and exported as “fastagr” and the potential cross-labeling or self-labeling probes were evaluated as allowed on the programme. The designed DNA probes will bind to

complementary 16S rRNA of peculiar bacterium or group of bacteria from purified and amplified 16S rRNA.

2.7 Probe Evaluation/Screening (Capillary Gel Electrophoresis)

The identified probes (biomers.net) were evaluated on selected samples from the clone library. Using 2 target and 3 non-target each of 6 probe, on a 36cm 3130xl capillary array (Applied Biosystems) in the ABI Genetic Analyzer 3130xl sequencer (Applied Biosystems) containing optimzed polyer 7 (POP-7) with injection time set at 16 – 22 seconds and electrophoretic conditions, run-time 180 seconds at 15 KV, current 100 μA and 60°C run temperature. Finally, 7 probes which includes the Universal probe, were eventually screened on the 75 children normalised DNA samples, 1 ng/μL in this study.

The evaluation and screening of the designed probe by capillary gel electrophoresis, started with prior flourecence end-labelling, on ice, of the specific DNA probes for detection. The end-labelling was achieved in the dark as far as possible by the patented GA "two step Single Nucleotide Extension (SNE)" reaction where the fluorescent dye TAMARA binds to ddCTP (5-proparglyamino-ddCTP-5/6 TAMARA). (Vebo et al., 2011) The reaction mix contained 0.25 U Hot Termipol DNA polymerase; 1X Hot Termipol buffer C; 4 mM $MgCl_2$; 0.4 μM 5-proparglyamino-ddCTP-5/6 TAMARA; 5 μL H_2O; 0.1 μM of each designed DNA probe and 2 μL 10X diluted ExoSAP treated 16S rRNA PCR products (approximately < 1 ng/μL) as earlier (direct sequencing). With reaction mix centrifuged at 3000 rpm for 1 minute and loaded on the pre set PCR thermocycler parameters; initial denaturation step, 95°C for 12 minutes; 5 cycles of 96°C for 20 seconds and 60°C for 30 seconds denaturation and annealing conditions respectively; and a final elongation step for 5 seconds at 60°C.

In addition, at 37°C and 80°C for 1 hour and 15 minutes respectively in the PCR thermocycler, the end-labelled reaction mix was incubated with 1 μL SAP treatment, removing any residual nucleotides and phosphate groups from the 5' end of the PCR amplified bacterial 16S rRNA gene. If not assayed immediately completed reaction mix was stored at 4° C for at least 24 hours or at -20°C for longer.

The reaction mix for capillary gel electrophoresis was 0.5 μL LIZ 120 (Applied Biosystem); 9 μL HiDi formamide and 1 μL SAP treated end-labeled product and incubated at 95°C for 5

minutes in the PCR thermocycler. The mix was centrifuged at 3000 rpm and loaded according to parameters previously used the sequencer earlier in the probe evaluation.

Generated signal peaks were scruitinised, base-called in the Gene Mapper 4.0 programme and similarly transformed numerically into peak "heights" (PH) and "area" (PA) for further statistical analysis. On the Gene Mapper 4.0 programme >size manage editor (used to set the standard for the peaks generated) < display plots (displayed table of heights and area of peaks)

Logistic Regression statistical analysis was performed on the SYSTAT programme using the PH and/or PA values obtained.

2.8 Bioinformatics

Forward and reverse sequence reaction results were assembled, aligned and trimmed for noise within the highly conserved region using the freely downloaded CLC Software from CLC bio A/S (http://www.clcbio.com/index.php?id=479) and/or the BioEdit Software (http://www.mbio.ncsu.edu/bioedit/bioedit.html).

The 16S rRNA gene sequences were further identified using the Ribosomal Database Project classifier for assigning phylogenetically consistent higher-order bacterial taxonomy (http://rdp.cme.msu.edu/classifier/classifier.jsp).

2.8.1 Chimeric Sequences

Empirically selected "high-quality" assembled and aligned sequences were filtered of chimeric 16S rRNA with the chimeric slayer algorithm in mothur (http://www.mothur.org/) prior to further analysis. "High-quality" sequence are thus described as sequences with less mismatch in the forward and reverse sequences consensus assembly and appropriately not short length. The relevant commands were used in the chimera slayer on the Mothur programme and as kindly assisted by Trine Frisli (PhD staff at HiHm).

Sequences in fasta format were uploaded onto the programme (Mothur). Details of complete commands used in this programme is as listed in the Appendix, *Figure* 2.8.1A. A summary of the input sequences was generated with a command. In the event of sequences with unexpected lengths, filtering was commanded. With appropriate commands, candidate and template (reference) files which have been uploaded are aligned and set to chimera removal

base on the chimera slayer algorithm. Finally, a summary of the identified unique sequences were commanded and a file name was created for the sequences in the collection.

This was adopted from KE Ashelford *et al*. (2006). The sequences that was screened for chimera were transformed into programme appropriate fasta format. Reference sequences devoid of chimera, "The rRNA 16S.gold.NAST_ALIGNED.fasta reference alignment " was used as standard and as stipulated in the mothur programme.

2.9 Statistical Analysis (including MCR analysis)

The large set of mixed sequences were kindly resolved by Professor Knut Rudi using the Multivariate Curve Resolution (MCR) analysis in an attempt as to expose and recover the pure components in the spectral of sequences. All the analyses of sequence spectra were performed using MATLAB® R2010a software (The MathWorks Inc., Natick MA, USA), Statistical and Bioinformatics toolboxes for MATLAB®. For EFA, PCA and MCR analyses, PLS Toolbox v5.8 for MATLAB® (Eigenvector Research Inc., USA) were used.

Firstly, an alignment of all of the mixed sequences spectra were generated and repeated for pre-processing and normalisation taking only small portions of individual peaks in the spectra to avoid peak shifts due to differences in retention times. Performed PCA and/or EFA determined the number of significant components explaining the most variations in the data set. With the pre-determined component number setting, MCR was run on the aligned spectra. MCR output are the information on the relative amount of each components in the individual sample/sequence in the data set and the based called spectra information on each components.

The detailed process and commands used in the resolution is as appended in *Figure* 2.9.1A (Appendix).

2.9.1 Regression analysis

In order to find any correlation in the data generated and disease diagnostics, regression analysis was performed. The was achieved using the SYSTAT Logistic Regression programme as per programme outlines (http://www.scientificcomputing.com/articles-DA-SYSTAT-13-110310.aspx) as revelant to our large data set and mostly recommended for clinical sampling studies. Binary logistic regression analysis tool was used in modeling, for

its specificity in regression diagnostics of the categorical dependent variables of disease (UC [3] or CD [2]) or not diseased (Control [1]) against independent continunious variable of either the peak area (PA) or peak height (PH) as extrapolated from the electrophoregram from capillary gel electrophoresis sequencing. From the main menu of the SYSTAT window, an output (outcome) was achieved by selecting in this specific order: Analyse; Regression; Logistic; and Binary, and within the appearing “dialog box” selected from the preloaded variables data with default parameters except that the “constant” was unchecked in the analysis of the direct sequencing MCR values (AV_COMP x).

3. Results

3.1 DNA Extraction/Purification

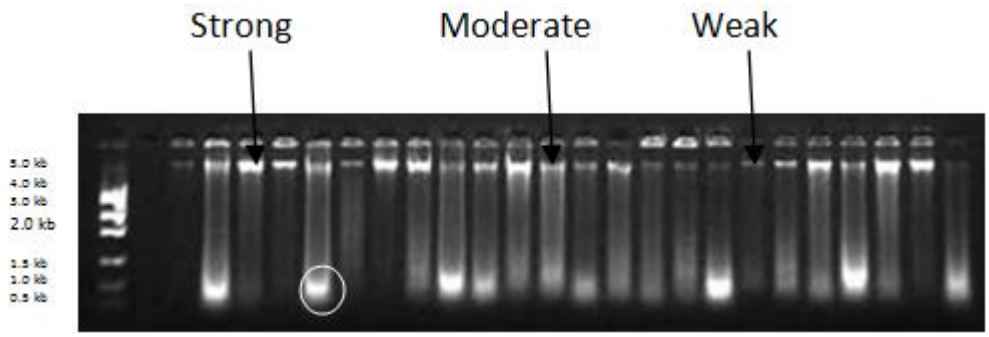

Figure 3.1.1: Post isolation DNA quantification. With a 1 kb DNA gel size ladder, feacal bacterial DNA was visualised at >5 kb. The DNA band strength was varied and was thus graded as indicated, Strong, Moderate and Weak. Sheared or degraded products are shown in circled lane and graded similarly.

DNA band visualised showed varying intensity and these strengths (Strong, Moderate or Weak) were noted as shown with images in *Figure* 4.1.1. The full list of all samples as graded are tabulated in *Figure* 3.1.1A (Appendix), DNA concentration and optical density (OD) ratio ($Ab_{260/280}$) showed no trends with DNA gel strength and observed degradation products or sheared DNA as visibly very thick bands towards the end on gel lane. The OD ratio was not consistent within the reference range of ~1.8 (pure DNA) and ~2.0 (pure RNA) which could not be explained as spectrophotometric light interference from sheared DNA or degraded products. However, they were still acceptable for the amplified "housekeeping" gene, 16S rRNA analysis as seen in electrophoresis gels confirmation prior to further analysis (not shown).

3.2 Direct Sequencing

a.) Sequence quality

The sequencing of the amplified bacterial 16S DNA pool of 75 childreen stool samples, overall produced nucleotides information of acceptable quality value. Specifically, sequence were qualifiied as saturated, strong, moderate, weak and poor, shown in *Figure* 3.2.1 with excerpts of electropherogram reads with grading in *Table* 4.2.1A, indicated that 60% and 25.33% of the sequenced nucleotides were deemed good and moderately good respectively

for analysis, according to ABI Biosystems handbook. About 1% of the nucleotides signal strength was saturated. Conversely, 8% were poor while 4% of the raw signal reads showed no readable nucleotide information.

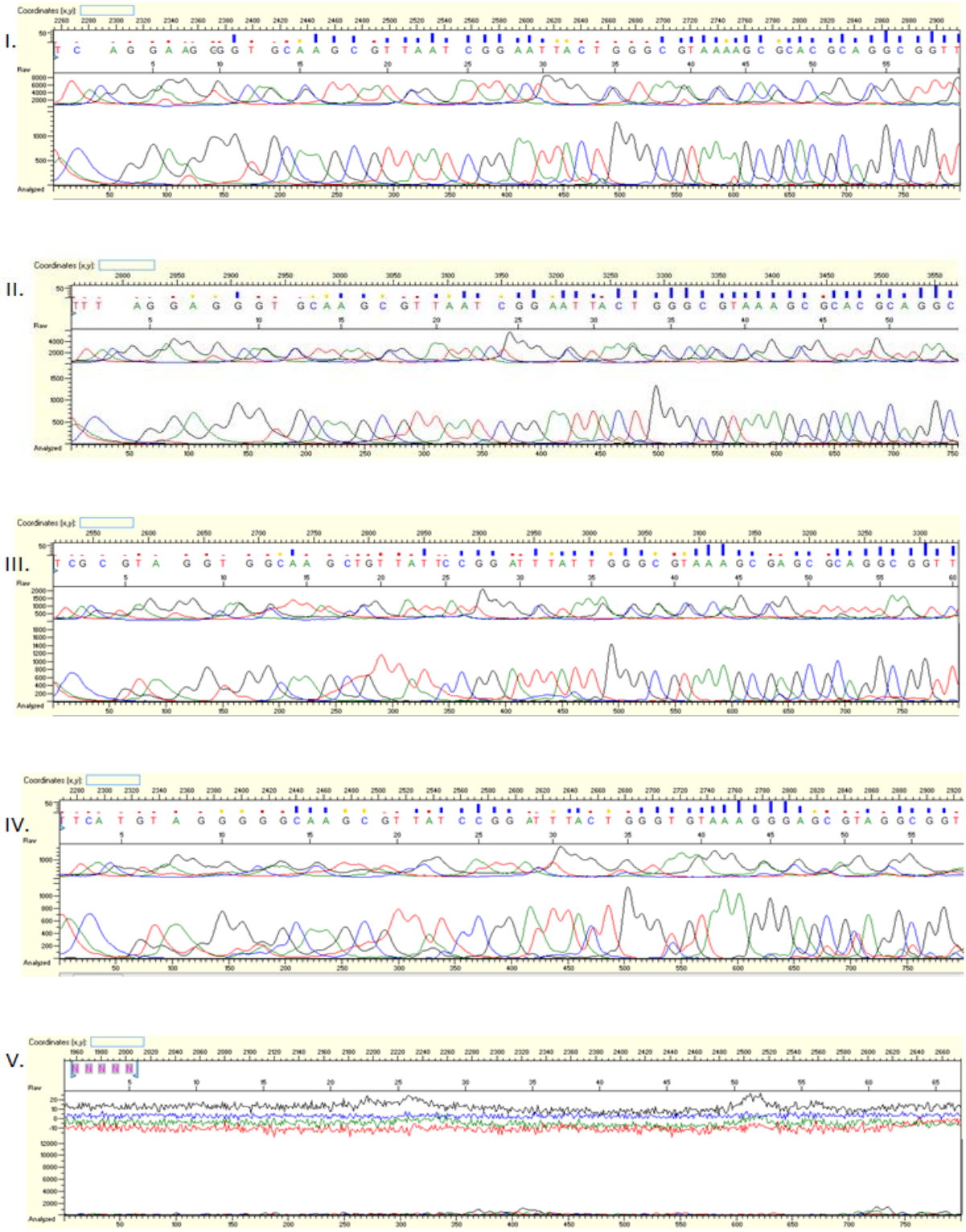

Figure 3.2.1: Sequence Electropherogram reads (excerpts) from the direct sequencing of the PCR amplified stool DNA by the automated ABI Genetic Analyzer 3130xl sequencer. Traces designated I, II, III, IV and V show grading of signal strength of the raw sequence quality as; Saturated (≥8000), Good (≥2000), Moderate (≥1000), Poor (≤1000) and Noise (No readable signals) respectively based on Raw trace (upper scan).

b.) Application of the Sequences (mixed) to Multivariate Curve Resolution

With no prior knowledge of the bacterial composition in each of the sampled pool of DNA sequenced, MCR treatment decomposed a concentration of 6 sequence components in the sequence spectral data. The *Table* 3.2.2 has enumerated the suggested MCR base called sequence components with the RDP search matches, *Faecalibacterium*, *Dialister*, *Haemophilus*, *Enterobacteriaceae* and *Bacteroides* as dominant within each pool scan. The base call of component 2 was a short read and thus was discarded as noise with no RDP sequence match revealed from search.

Table 3.2.2: MCR components from mixed sequences of 75 children different samples. Sequence Match from RDP database indicates abundant bacteria group in the pool of diverse bacterial sequences.

Components (match)	**Sequences**
Comp1 (*Faecalibacterium*)	AGCGTGTCCGATTTACTGGGTGTAAGGGAGCGCTAGCGGAGAGCAAGTTCGGAGTGAA ATCCATGGGCTCAACCCATGGAARTGCTTTCAAACCTTGMTTTTCTTTGYTAGTGCAAAG GTAAAGTCGGATRCCTGAGGTGGTACGGGTGGAATGCGTAATATTYGGAGGAACACCAT GGCAAGGCGGTCRTACTGGGCACCAACTGACGRTGAGGCTCAA
Comp2 (Designated Noise)	AGCTAGTATCCGGATTCTARTGGGTGTAAAGGGCGTAGCGGTTATCTAAAGGGCTTTT
Comp3 (*Dialister*)	AGCGTTGTCCGGATTATTGGGCGTAAAGCGCGCGCAGGCGGCTTTCCRAAGTCCTCTCTT AAAAGTGCGGGGCTTAACCCCMGYTG GGGYATGYAACCTGGYAAYCCTGGAGTATCGGAYAGYAAAGMGAGAATTCCATAGTGT AGCGGTYAAATGCGTAAGATTAGGAAG AACACCGGTGGCGAAGGSGACTTTCTGGACAAAACTGACGCTGAGGCGCGAAA
Comp4 (*Haemophilus*)	AGCGTTATTCGGAATAARTGGGCGTAAAGGGCACGCAGGCGGTGKCTTAAGTGAGGTGT GAAAGCGCCCGGGCTTAACCTGGGAATMGCATTTCATACTGGGGTGCCGTAAMTACTTT AGGGAGYGGTAYATATTCTCACGTMGTAGCGGTYAAWGTSCTTAAGTATGTGAAGYAA TACCGAAGGCAGAAGSCARCCCTMGGAWTGTCACGTGACSRTCATGTGCAAA
Comp5 (*Enterobacteriaceae*)	AGCGTTAATCGGAATTACTGGGCGTAAAGCGCACGCGGCGGTTTGTTAAGTCAGATGTG AAATCCCCGGGCTCAACCTGGGAACTGCATCTGATACTGGCAAGCTTGAGTCTCGTAGA GGGGGGTAGAATTCCAGGTGTAGCGGTGAAATGCGTAGAGATCTGGAGGAATACCGGT GGCGAAGGCGGCCCCCTGGACGAAGACTGACGCTCAGGTGCGAAA
Comp6 (*Bacteroides*)	ACGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGTGGARTTGTTAAGTCATGTATGT GAAAGCTTTGCGGCTCAACCGTAAAATTGCATTTGAWACTGGAAGWCTTGAGTGCAGT AGAGGRAGAGGCGGAATTCCTGGTGTAGCGGTGAAATGCTAATATCACGAAGAACATC CGATGTGCGAAGGCGGCTTAGCTGGACTGTAACTGACYRTGAMGCTCGAAA

Consequently, the MCR technique recovered numerical data as average component frequency values for each of the sampled sequences. The highest component values within each sample DNA sequence formed the bases for the selection of DNA sample for cloning. In the bar chart, *Figure* 3.2.3, cloned selection is ploted and the overall data can be perused in the Appendix, *Figure* 3.3.3A.

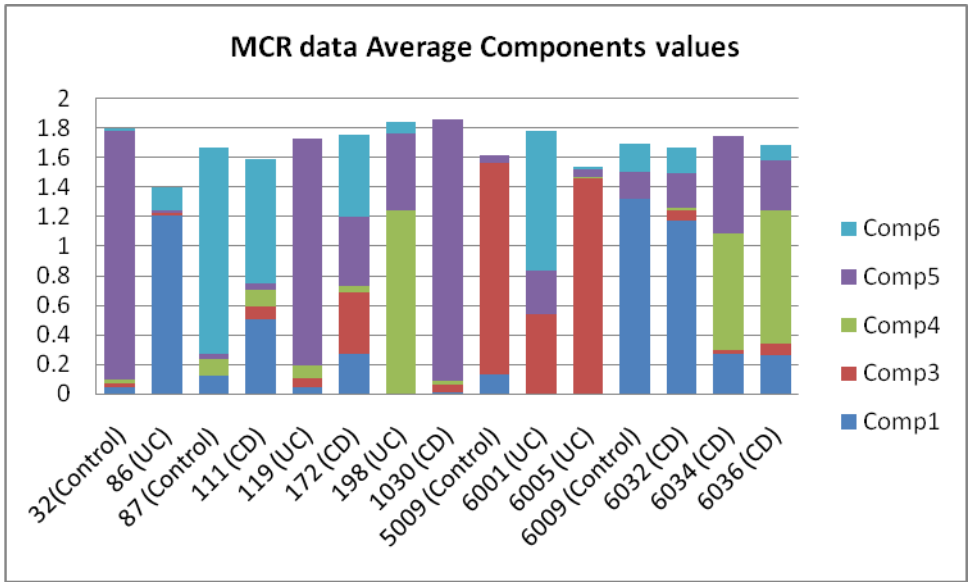

Figure 3.2.3: MCR data as used for clone library collection. Bar Chart value represents average frequency of each component (as colour codings) in the resolved mixed sequences. Within this selection, the most abundant (highest) MCR average component (1-6) value for each diagnosis (Control,CD and UC) was assumed for cloning [Samples 32, 86, 87, 111, 119, 172, 198, 1030, 5009, 6001, 6005, 6009, 6032, 6034 and 6036].

c.) Relating sequences to MCR data

An overrepresentation of the *Enterobacteriaceae* (Comp5) was indicated in the mixed sequence-MCR component abundance analysis matrix (*Figure* 3.2.4) and closely followed in abundance was the *Faecalibacterium* (Comp1) and *Bacteroides* (Comp6) in the sampled 16S rRNA sequences. *Haempohilus* (Comp4) and *Dialister* (Comp3) with high intensity of blue suggest underrepresentation within the sampled sequence pool. In other words, samples cloned (as indicated with the arrows in *Figure* 3.2.4) were representative of the samples with relatively overrepresented components as to extrapolate any correlation with disease.

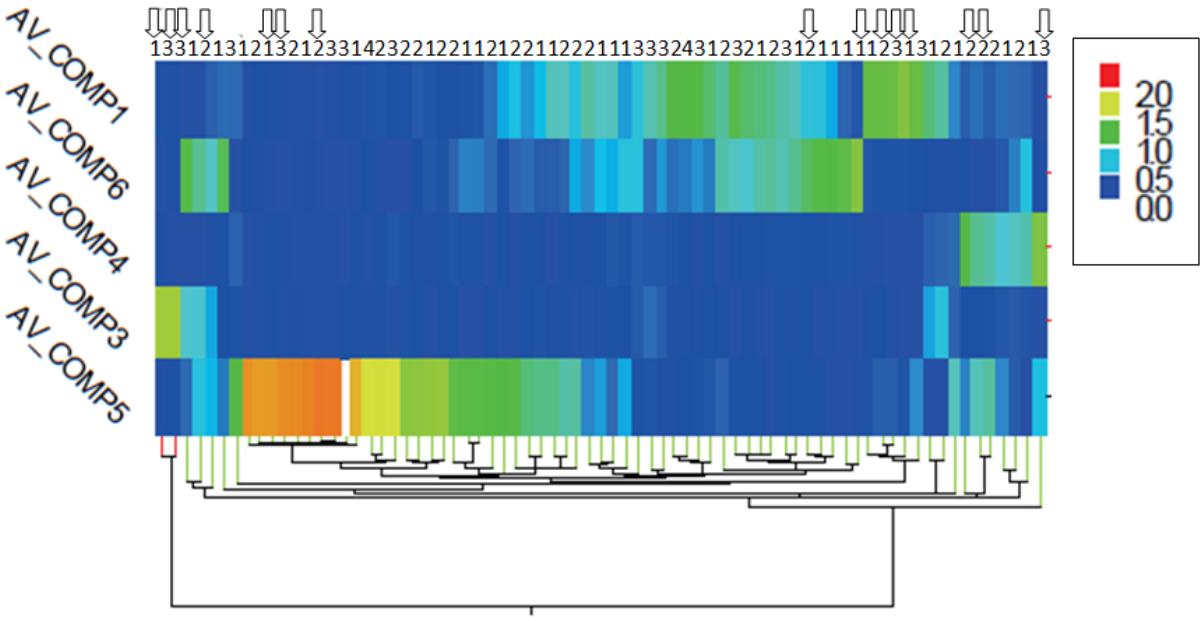

Figure 4.2.4: A permuted matrix relating the calculated MCR average component (Av. Comp 1, 3, 4, 5 and 6) abundance within each specimen faecal DNA sequenced. Selected samples that were cloned are indicated with the pointing arrow. Diagnosis represented for each sample; 1=Control, 2=CD; 3=UC and 4=IBDU. [Av. Comp ranges from 0.0 -2.0 with colour codes, Blue indicates underrepresentation of components, while red indicates an overrepresentation of components for each sampled DNA sequences.Conversely, colours inbetween a representative abundance of components].

Also, as indicated in the permuted matrix (*Figure* 3.2.4), there seem a taxanomic relationship among the species communities in the DNA sequence pool except for two samples, one each of control and UC, with a different lineage from the rest.

3.3 Cloning results

Frequency of transformants with inserts analysed by colony PCR as shown in gel picture of *Figure* 3.3.1, gave a 52.1% (100/192) for chemically transformed bacterial 16S rRNA fragments (~1400bp) and in most cases, incubation time of transformed fragments lasted several days to generate visible colonies. However, the greater efficiency of transformation was achieved by the electroporation protocol which raised frequency to 76% (513/675) from overnight incubation.

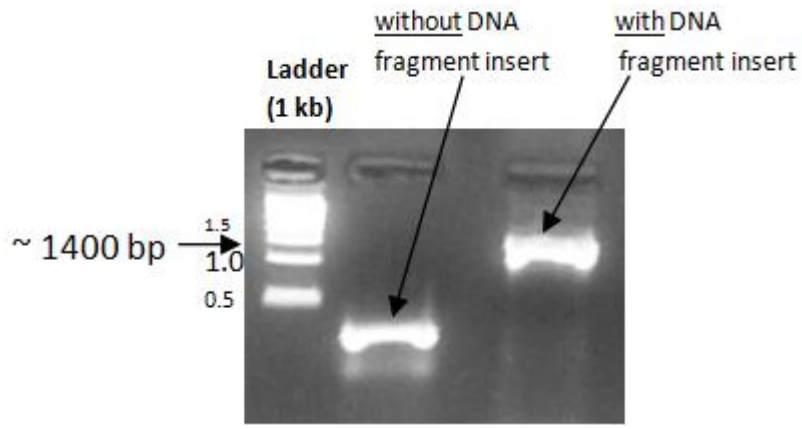

Figure 3.3.1: Colony PCR gel electrophoresis analysis of clone transformants. PCR product of picked clones (white or light blue), analysed by 1.5% agarose gel electrophoresis showed lanes with or without inserts of DNA fragment, transformed using TOPO® Vector in One Shot® Competent Cells (*E. coli*) by electroporation or chemical cloning. Frequency of positive transformant (clones with inserts) analysed was greater by electroporation as compared to chemical cloning.

a.) Clone library

From MCR data values as in excerpts ploted in the bar chart above, 15 16S DNA samples were selectively cloned, generating about 600 clone sequences. With 308 assembled (forward/reverse) sequences of appreciable quality, 283 sequence were devoid of chimeria from the chimera slayer algorithm and thus formed the clone library for any further analysis. As shown in the clone library is representative of human gut bacterial load.

Table 3.3.2: RDP 16S rRNA sequences matches summary for clone library sequence. Phylotype abundance given in brackets for each diagnosis.

Domain **(Control, CD, UC)**	**Phylum** **(Control, CD, UC)**	**Genus** **(Control, CD, UC)**
Bacteria (49, 112, 122)	Firmicutes (23, 59, 49)	Dialister (5, 1, 16)
		Enterococcus (0, 14, 0)
		Faecalibacterium (3, 1, 9)
		Lactobacillus (0, 1, 19)
		Subdoligranulum (2, 17, 3)
		Unclassified genus (3, 7, 8)
	Proteobacteria (20, 39, 48)	E.coli (20, 19, 23)
		Haemophilus (0, 16, 18)
		Unclassified Pasteurellaceae (0, 3,7)
	Bacteroidetes (6, 12, 24)	Bacteroides (2, 8, 22)
	Verrucomicrobia (0, 1, 0)	Akkermansia (0, 1, 0)
	TM7 (0, 1, 0)	unclassified

An RDP search for sequence matches in the clone library, retrieved the listed phylum and genus in the *Table* 3.3.2 above with the complete list in Appendix (*Table* 3.3.2A). Sequences reads recovered from the sequenced clones suggest the Firmicutes and Proteobacteria were the most dominant phyla and with abundance within the diseased (CD or UC) category. However, there was also a shift in most phylotype abundance tendency within the Firmicutes

in CD cases (59) as opposed to UC cases (48) in Proteobacteria. In other words, the phylotypes that was determined to be increased in UC (Proteobacteria and Bacteroidetes) as compared greater CD phylotypes in Firmicutes. Bacteriodetes are also seen to be deplected in comparison to the Firmicutes and Proteobacteria groups. However, within each phylotype recovered, there were raised in the diseased category in comparison to the control group except for Verrucomicrobia and TM7.

Haemophilus RDP matches from clone sequences was in the disease category, CD and UC, with almost same abundance 16 and 18 respectively and not at all from clone sequences in the control groups. This was noted with an interest as this has not been previously reported.

b.) Phylogeny

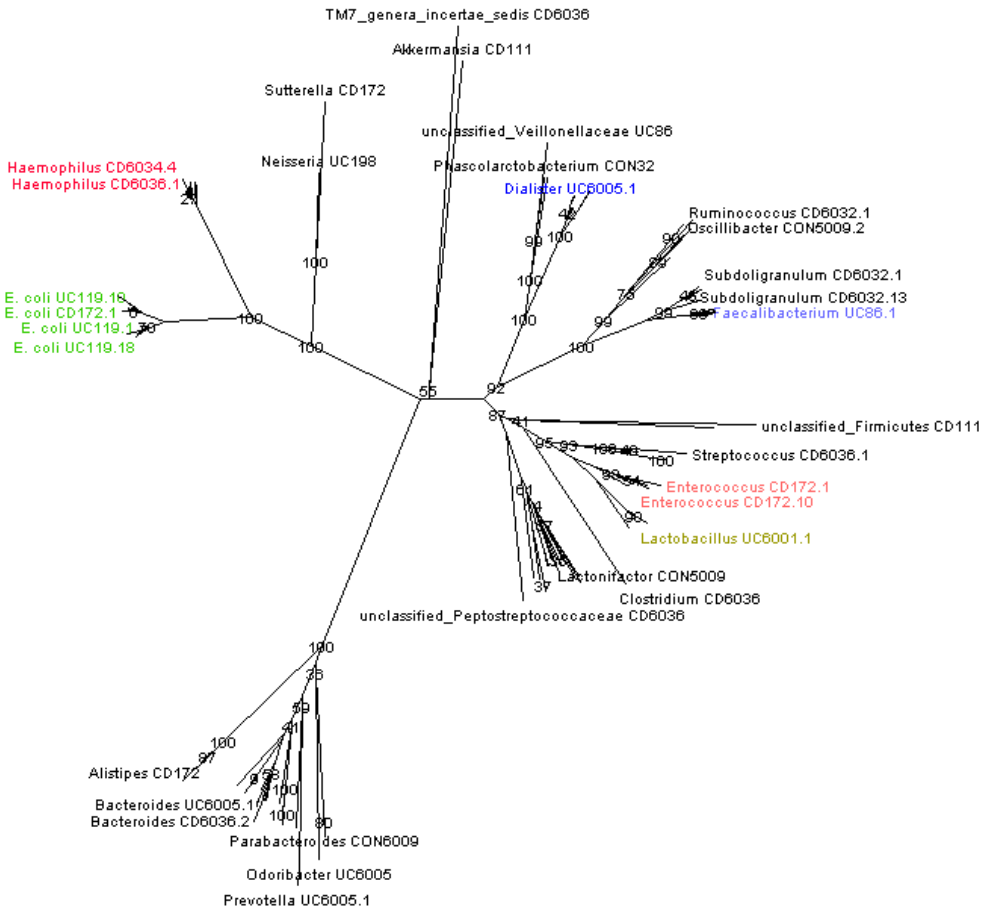

Figure 3.3.3: Phylogenetic tree from CLC genomic tool. A radial phylogram illustration of the taxanomic relationships of the clone library sequences (RDP match) of the 16S rRNA genes as a multiple tree drawn using the Dendroscope programme. Clone identity (e.g 6001.1) and diagnosis (Control,CD & UC) indicated with root/node distances along trees. Similarly, coloured bacteria species (*Haemophilus*, *E. coli*, *Dialister*, *Faecalibacterium*, *Enterococcus*, and *Lactobacillus*) indictate coverage within the tree of the six designed probes screened for any correlation to disease (IBD) diagnostics.

The dendrogram illustrated in *Figure* 3.3.3 is a representation of the phylogeny and diversity of bacterial sequences in the clone library of the samples selectively cloned in a cladistic tree. Basically, four different clusters of bacteria are represent from the sequence library. This includes the three phyla; gram positive Firmicutes (*Subdoligranulum*, *Faecalibacterium*, *Clostridium*, *Ruminococcus*, *Dialister*); facultative anaerobic species Proteobacteria (*E.coli*) and gram negative Bacteroidetes (*Parabacteroides*, *Bacteroides*) and the other groups includes TM7, a subgroup of gram-positive bacteria.

From the clone library sequence matches, only *Faecalibacterium* and *Dialister* were matched in the MCR components resolved.

3.4 Probe identification

a.) PCA Clusters

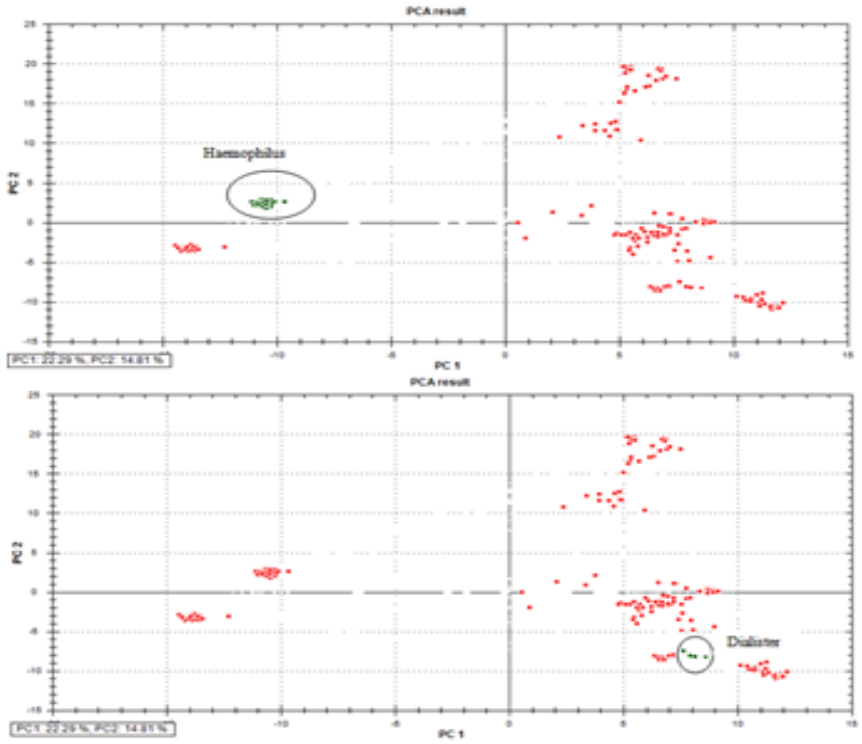

Figure 3.4.1: PCA plot indicating 16S rRNA sequence clusters for probe design. Targets (*Haemophilus*, top or *Dialister*, bottom) indicated in defined clusters (green) and excluded non-target (red). From a total of 303 forward/reverse assembled 16S sequences clone libray of children gut bacteria, 282 chimera free sequences generated over 19 probes according to PCA clustering of related 16S sequences.

In the design of the probes, clustering of all the 283 clone library devoid of chimeras are as represented in the quadrants of the PCA plot in *Figure* 3.4.1. Within each window (PCA1,PCA2,PCA3….PCAx) as shown in the axis, probes sequences generated in the GA "in-house" TNTprobe programme revealed quite a lot of probe sequences, of which 19 probes (list in *Table* 3.4.1A in the Appendix) were quality enough in theory, as verified by the programme the used to hybrise on targets and exclude non-target. In these windows, clusters reflected sequence similarities and probes identified sequences targets such *Dialister* and *Haemphilus* (indicated with circles in *Figure* 3.4.1) from the clone library sequences.

b.) Probe Evaluation/Screening

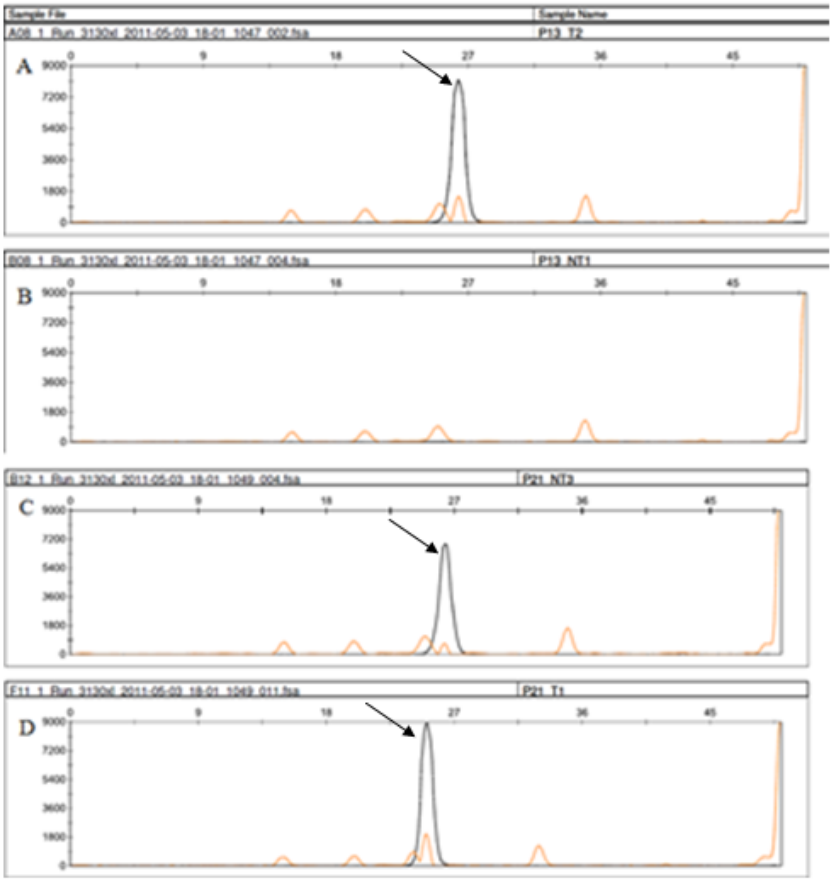

Figure 3.4.2: Signal reads from Probe Evaluation by capillary gel electrophoresis. Probe treated samples illicits hybridisation of target (A and D) sequences as peak (shown with pointing arrow) and excluded non-targets (B and C). However, a labeled peak in both targets and non-targets (C and D), eliminates P21 (Probe 21, targeting *Faecalibacterium*) in the above figure as a choice screening probe for all 16S rRNA sequence from the children samples.

Probes evaluated that showed specificity were fit for screening on all PCR amplified 16S rRNA gene samples in this study. There were 6 probes screened as well the Universal probe

(*Table* 3.4.3). Probes that are labeled on its target as peaks as indicated with the pointing arrows in *Figure* 3.4.2 were deemed successful. Thus the size of the peak gave an indication of the abundance of the bacterial target that the probe labelled. The successfully labeled samples from the capillary gel electropherograms reads were extrapolated as "peak heights" (PA) (See *Table* 3.4.2A in Appendix) on the GeneMapper programme for statistical analysis.

Table 3.4.3: List of Probes, sequences and the targeted bacterial in brackets.

Probe # (Targets)	Probe Sequences
P_3 *(E. coli)*	GCCTCAAGGGCACAAC
P_6 *(Dialister)*	AAGAACTCCGCATTTCTGC
P_8 *(Faecalibacterium)*	CGTAGTTAGCCGTCACTTC
P_13 *(Haemophilus)*	TCGCTTCCCTCTGTATACG
P_16 *(Enterococcus)*	CCCTCCAACACTTAGCA
P_18 *(Lactobacillus)*	CCTGTTTGCTACCCATACTTT
UNI01 (16S Universal)	CGTATTACCGCGGCTGCTGGCA

3.5 Logistics Regression Analysis

The "peak heights" numerical values earlier as extrapolated and analysed by SYSTAT binary logistics regression for its unique and linear output application in disease diagnostics. This exposed any implication of probes to diseases diagnostics in relation to the non-IBD control subjects. Similar analysis was performed with the MCR component frequency values.

a.) MCR data

Table 3.5.1: Binary Logistic Regression analysis on MCR data. Significance (p <0.05) is highlighted ** in targets *Enterobacteriaceae* (UC) and *Faecalibacterium* (IBD) subjects with negative (-) and positive (+) correlation estimates to diseases respectively, in relation to control cases in the study. *Faecalibacterium* (CD) value highlighted with * (p= 0.05 – 0.1) showed some positive (+) correlation estimates to disease in relation to control but not significant.

Components	***p*-values** (Estimates Direction,+/-)		
	IBD	**CD**	**UC**
AV_COMP1 (*Faecalibacterium*)	0.036** (+)	0.096* (+)	0.129 (+)
AV_COMP3 (*Dialister*)	0.324 (-)	0.595 (-)	0.102 (-)
AV_COMP4 (*Haemophilus*)	0.747 (+)	0.443 (-)	0.165 (+)
AV_COMP5 (*Enterobacteraceae*)	0.626 (+)	0.485 (+)	0.049** (-)
AV_COMP6 (*Bacteroides*)	0.254 (-)	0.113 (-)	0.201 (-)

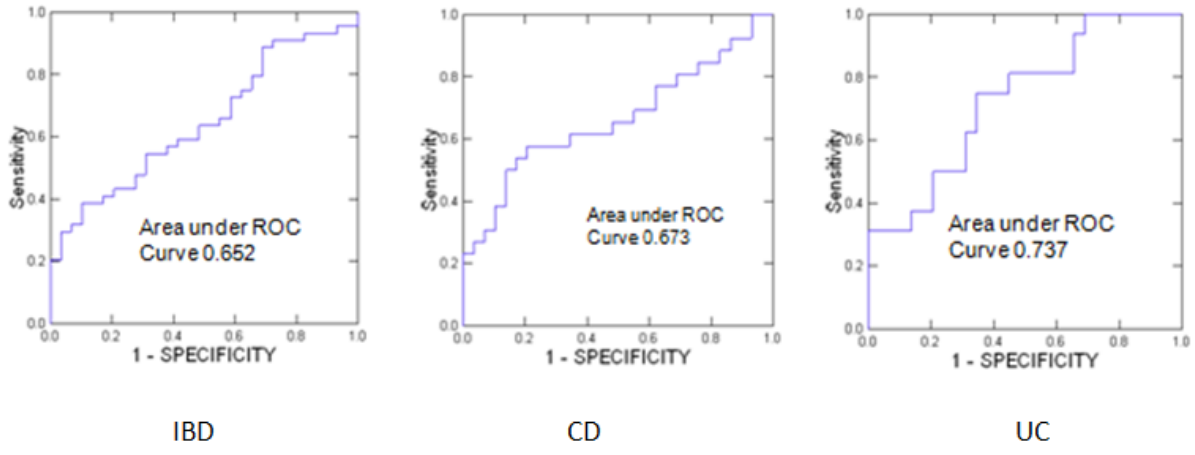

Figure 3.5.1: Binary Logistic Regression analysis on MCR data. ROC curve characterisation of the IBD data in the model is specific but lowest in sensitivity. Area under ROC curve is acceptable (≥ 0.5).

In the MCR data, *Faecalibacterium* (Comp1, p=0.036) and *Enterobacteroceae* (Comp5, p=0.49) were statistically significant (*Table* 4.5.1), regression estimates correlating positively and negatively respectively as potential candidates in IBD and UC diagnotics in the samples under studied versus control samples. Although there was a positive correlation

for *Faecalibacterium* in CD subjects but not statistically significant (Comp1, p=0.096). *Figure* 4.5.1 gives an overview of the fit of the statistical model used, which was acceptable for each diagnosis (ROC curve > 0.5).

b.) Probe Screening data

Similar analysis as in MCR data, performed on probe peak height (PH) values, normalised with universal probes, significantly implicated *Haemophilus* (Probe_13, p=0.028) and *Dialister* (Probe_6, p= 0.058) with negative and positive correlations estimates respectively in the labelled samples in CD and UC (*Table* 3.5.2). The constant regression estimates (p= 0.031) as shown, was significant and negatively correlating to UC, is suggestive of an unexplained factor in the model. As in the MCR data, the fit of the model was deemed satisfactory (ROC curve > 0.5).

Table 3.5.2: Binary Logistics Regression analysis of Probe Data. Positive (+) estimated correlation for both diseases CD (*Haemophilus*) and UC (*Dialister*) in relation to control data. Probe significant (p<0.05) for each disease is highlighted** (marginally significant* for *Haemphilus* [IBD]).

Probes	***p*-values** (Estimates Direction, +/-)		
	IBD	**CD**	**UC**
NP_3H (*E. coli*)	0.4 (+)	0.761 (+)	0.437 (+)
NP_6H (*Dialister*)	0.405 (+)	0.993 (+)	0.058* (+)
NP_8H (*Faecalibacterium*)	0.398 (-)	0.919 (+)	0.212 (-)
NP_13H (*Haemophilus*)	0.078* (-)	0.028** (-)	0.227 (+)
NP_16H (*Enterococcus*)	0.471 (+)	0.194 (+)	0.514 (+)
NP_18H (*Lactobacillus*)	0.189 (+)	0.174 (+)	0.1 (+)
Constant	0.837 (-)	0.458 (-)	0.031** (-)

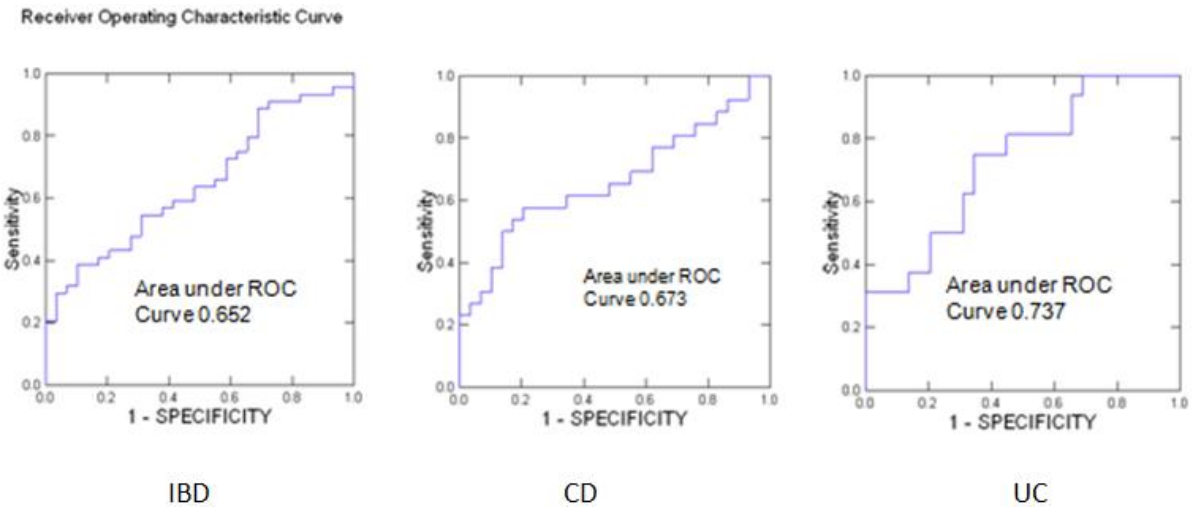

Figure 3.5.2: Binary Logistics Regression analysis of Probe Data. ROC curve expresses the quality of data analysed in terms of specificity and sensitivity (ROC curve > 0.5) in the model used

4. Discussion

4.1 DNA purity

The purity of DNA preparation is essential for sensitive and useful biological analysis. The evaluation of quality in the extracted bacterial DNA from the children samples, showed some inconsistences in concentration within reference ranges of pure DNA (≥ ~1.8 or ≤ ~2.0) and can be otherwise expected from the complex human faecal matter. (Li et al., 2003) Degraded products visualised at the bottom of the gel column may have affected this measurement. It may as be speculated that the repeated thawing and freezing of this sample over time can be responsible for any effect of DNA quality. (Li, et al., 2003) However, these degraded products could be RNA or sheared DNA but treatment with RNAse may have excluded RNA. Similary, despite using the highly comparable commercial kits, QIAamp® DNA Stool Mini Kit (Li, et al., 2003; McOrist, et al., 2002), for the extraction and DNA purification method, results obtained from the extraction and purification of the human faecal bacterial DNA in the samples under study here, may have reflected individual differences in extraction efficacy as there was no trend of uniform effect of the degraded products on concentration. (McOrist, et al., 2002)

The microbiota profiling analysis utilised the PCR amplified 16S rRNA gene as the established genetic marker for microbial community profiling, compared to previously reported studies by Li M *et al* (2003) as appropriately effective in the profiling of the gut microbiota. Therefore, the extraction method produced DNA eluates devoid of PCR inhibitors (McOrist, et al., 2002). In other words, the degraded products are no impediment to such analysis utilising the 16S gene as has been demonstrated in results presented in this study.

4.2 MCR prediction of abundant bacterial sequences (species) in cohort

The rigors of finding structure in an environment of mixed population was initially possible with the proven multivariate technique which is applicable to gut bacterial community (K. Rudi, et al., 2007; Trosvik, et al., 2007) as in the diverse bacterial sequences generated in mixed DNA pool in the cohort. By doing this, were able to explore selected DNA samples further, and the study on bacterial diversity and richness can thus be enhanced on other

samples in the study. As demonstrated by Trosvik P, et al (2007), their method that applys MCR showed that determing species of high-richness among community members may be a potential disadvantage of MCR application to prediction of total abundance in a data set but nonetheless can expose subgroups within a complex data set. Hence, to truly achieve the aim of mining the gut microbiota and establish biodiversity for diagnostic potential, the understanding of the relevant and representative microbial species is essential in the large data sets that are analysed in microbial studies. (K. Rudi, et al., 2007)

Generally, a consensus on the dynamics of intestinal microbiota imbalances has been difficult owing to differences in approach, specimens used and conflicting findings in literature. However, from microbial 16S sequence resolution, a prediction of most bacterial representation in the cohort sequence data was reported. Enterobacteriaeceae representation in most of the mixed samples sequence in this cohort study mirror that found in faecal analysis with increased concentration of enterobacteria in CD, which have implication in intestinal virulence (Sokol, Lay, et al., 2008) and in some other cases represented as *Enterobacter cowanii*, a relative within the family (Mondot, et al., 2011). Hence, it is not a coincidence predicting the increased abundance expressed in our samples. Similarly, an increased abundance of *Faecalibacterium* and *Bacteroides* was also observed in overall sequences-MCR components comparism matrix as well. However, contrary to this MCR prediction, this groups are thought to be deplected in published works from CD studies in relation to control subjects. (Joossens et al., 2011; Sokol, Lay, et al., 2008; Sokol, Pigneur, et al., 2008) The least representation of *Dialister* and *Haempohilus* components in these sequences is noted with the former decreased in CD patients (Joossens, et al., 2011; Sokol, Pigneur, et al., 2008) while the later exposed a previously uncharacterised gut phylotype in literature.

Importantly, these suggestions from MCR prediction are based on structure within the mixed sequences. In this view, further work such as the probe assay results can establish its use in IBD diagnostic.

4.3 Clone library evaluation of MCR prediction

There are basically three groups of highly abundant bacterial phyla represented in the selected children faecal DNA sequenced from cloning with frequently reported dysbiosis suggesting implication in the pathogenesis of IBD. Submissions here in relation clone library

results is deemed a representation of a subset of the cohort data based on MCR selected samples for cloning and the pure sequences determined. As a matter of fact, they showed biodiversity as achieved by previous study establish microbiota species by comparison to healthy controls from the cloning of bacterial 16S genes from six faecal samples but however, a larger library (25 000 clones) was generated for both sets. (Manichanh et al., 2006)

Within the clone library, retrieved sequences as matched included an array of commensal bacteria subgroups, *Dialister, E. coli*, *Faecalibacterium* and *Subdoligranulum*, which have otherwise been previously reported in literature to be implicated in gut bacteria dysbiosis between IBD and healthy subjects. (Joossens, et al., 2011; Mondot, et al., 2011; Sokol, Pigneur, et al., 2008) Interestingly, *Haemophilus* which has not been implicated previously was not infered from control cases in the children studied but represented in the diseased subjects. As such may have to be explored further for potential in diagnostic development. A small fraction of the library also harbour some "unclassified" bacterial group of the phylum Firmicutes. The biodiversity and reduced abundance of subgroups within the Firmicutes phylum have been interpreted in many studies but in some inclined to dorminance in CD subjects and this cohort clone library reflects Firmicutes richness in IBD cases. (Sokol, Pigneur, et al., 2008) Although biodiversity within Bacteroidetes phylum seem conserved (Sokol, Lay, et al., 2008), and the library in these study showed that there is an increase in the Bacteroidetes species abundance versus the control especially in CD (Manichanh, et al., 2006).

In the view that the clone library is to reflect the cloning sample selected due to its high MCR scores in the predict components. The identification of most of these components from the clone library very much justifies its selection criteria in such a way as to establish *Enterococcus* as the dorminant group of the Enterobacteriaceae family identified in the mixed sequence resolution. This should be somewhat increased in IBD cases especially CD although some reports has stipulated a decrease (Frank et al., 2007; Mondot, et al., 2011; Sokol, Lay, et al., 2008), and this happened to be significantly correlating to UC instead by our binary logistic model in relation to control. Similary, *Faecalibacterium* predicted in the MCR analysis was equally significant by the same model in IBD cases and only marginally correlating to CD, which is also antagonistic to previous reports by Sokol *et al* (2008).

It would also be worthy to note that the unculturable TM7 phylotypes which are usually found in natural environmental habitats and now associated with higher diversity in CD than in controls and UC, only recorded one representation in CD clones matched. However, would have expected much in this cohort given that the samples collected were from early and untreated cases of IBD where TM7 phylotypes are suggested to be community modulator or trigger of inflammation in the early stages of IBD. (Kuehbacher, et al., 2008)

It will also be interesting to point out that certain bacteria species which has been previously reported in faecal populations were missing in the clone library matches, such as *Bifidobacteria* which is demonstrated to be found in 80% of cultivable feacal bacteria with probitic potential (Picard et al., 2005) and an also equally gut beneficial *Lactobacillus* was predominant in UC and absent or almost absent in controls and CD. (Sokol, Lay, et al., 2008) Control samples may be non-IBD diagnosed but could as include IBS (Inflammatory Bowel Syndrome) cases, with similarities in symptoms to IBD. Similarly, this subjects may as been IBD without full manifestation in diagnosis. Essentially, in most reported studies where clone libraries have been generated from samples, they generally involved subjects undergoing or have undergone treatment which is different in this cohort of untreated and early diagnosed patients. (Sokol, Lay, et al., 2008)

4.4 Correlation in the MCR data and Probe screening

With respect to the MCR prediction, analysis implied significance in two bacteria groups, *Faecalibacterium* (IBD) and *Enterobacteriaceae* (UC). However, the probe data analysed, rather exposed two different bacterial species significantly correlating to CD (*Haemophilus*) and UC (*Dialister*) but within the five resolved species predicted by MCR. In order words the probe data regression showed the specificity of the target probes in IBD disease screening and its relevance to diagnostics. It is evident that this deduction from the probe analysis of this two species, reflects the abundance profile as compared to the controls, from the clone library of the representative samples. This would not be suprising and validates the probes that were designed with the sequences in the clone library as templates.

As time could permit, few probes were evaluated and even fewer screened, it may be said that the narrow results maybe impacting on confirmating the actual biodiversity in this cohort. Even though, a probe targeting the *E. coli* species group was screen, its worthy of note that this highly report bacteria group (Manichanh, et al., 2006; Mondot, et al., 2011;

Sokol, Lay, et al., 2008) had no significance in this reported data and not predicted by MCR. Conversely, within the probe data statistics model, there remains a significant constant of unexplained factor in UC which may be driving the results output. It would thus enhance confirmation and conclusion to explore effects such as age within the data set. Although, this was evaluated with potential but not as yet reported as results were inconclusive.

4.5 Further work

In order to further investigate the biodiversity of the gut microbiota and implication to IBD against control subjects it would be a suggestion to recruit even more samples with equal number of diagnosed subjects, that is for example; 100 controls, 100 UC and 100 CD patients in the cohort. Conversely, a randomly selected control group, instead of non-IBD controls with IBD symptoms reported in this cohort. In this way, any diagnostic potential could be validated and probably inconsistency better understood if not completely avoided. The clone library can as well be linked more to the MCR components by generating a Local Blast database library and search for component frequency in each of the cloned sequences.

As there were more probes designed than could be evaluated or screened and potential more windows to consider in the probe design to generate even more probes. An assay of a greater number of more 16S probes will enhance the realisation of an extensive enteric gut biomarker for IBD. Similary, with larger probe set, a microarray platform could be construsted and simulateneously screened for any correlations to the disease. In addition, the Luminex techonology as has been used in microbial community detection and biomarkers screening, is an ideal platform for clinical validation for the development of companion diagnostics.

Given that the samples analysed were patients recruited with early IBD symptoms and it would also be interesting to analyse the gut microbial diversity of same individuals post treatment. This should be particularly of interest to the management of the disease as is the confirmation of the early diagnostic potential of IBD from the aim of this study.

In another hand, other statistical approach and model could be used, such as analysis of variance (ANOVA) to establish disease correlation against the control group.

5. Conclusion

The evaluation of the children faecal samples highlighted the biodiversity of the gut microbiota. Similarly MCR predicted bacterial species and probes specific to bacterial species that could be targeted for potential addition to the already novel GA screening tools aimed at establishing the IBD diagnostic platform. The selective targets are Dialister and Haemophilus, as implicated in UC and CD respectively. This may have wider implication for early IBD diagnostic if followed up with extensive validation.

6. References

Achtman, M., & Wagner, M. (2008). Microbial diversity and the genetic nature of microbial species. [10.1038/nrmicro1872]. *Nat Rev Micro, 6*(6), 431-440.

Baumgart, D. C., & Sandborn, W. J. (2007). Inflammatory bowel disease: clinical aspects and established and evolving therapies. *The Lancet, 369*(9573), 1641-1657. doi: 10.1016/s0140-6736(07)60751-x

Blaser, M. J., & Falkow, S. (2009). What are the consequences of the disappearing human microbiota? [10.1038/nrmicro2245]. *Nat Rev Micro, 7*(12), 887-894.

Bouma, G., & Strober, W. (2003). The immunological and genetic basis of inflammatory bowel disease. [10.1038/nri1132]. *Nat Rev Immunol, 3*(7), 521-533.

Braun, J., & Wei, B. (2007). Body Traffic: Ecology, Genetics, and Immunity in Inflammatory Bowel Disease. *Annual Review of Pathology: Mechanisms of Disease, 2*(1), 401-429. doi: doi:10.1146/annurev.pathol.1.110304.100128

Braus, N. A., & Elliott, D. E. (2009). Advances in the pathogenesis and treatment of IBD. *Clinical Immunology, 132*(1), 1-9. doi: 10.1016/j.clim.2009.02.006

Brugère, J.-F., Mihajlovski, A., Missaoui, M., & Peyret, P. (2009). Tools for stools: the challenge of assessing human intestinal microbiota using molecular diagnostics. *Expert Review of Molecular Diagnostics, 9*(4), 353-365. doi: doi:10.1586/erm.09.16

Cohen, A. S., Najarian, D. R., Paulus, A., Guttman, A., Smith, J. A., & Karger, B. L. (1988). Rapid separation and purification of oligonucleotides by high-performance capillary gel electrophoresis. *Proceedings of the National Academy of Sciences, 85*(24), 9660-9663.

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., . . . Tiedje, J. M. (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Research, 37*(suppl 1), D141-D145. doi: 10.1093/nar/gkn879

Dethlefsen, L., McFall-Ngai, M., & Relman, D. A. (2007). An ecological and evolutionary perspective on human–microbe mutualism and disease. [Article]. *Nature, 449*(7164), 811-818. doi: 10.1038/nature06245

Frank, D. N., St. Amand, A. L., Feldman, R. A., Boedeker, E. C., Harpaz, N., & Pace, N. R. (2007). Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proceedings of the National Academy of Sciences, 104*(34), 13780-13785. doi: 10.1073/pnas.0706625104

Freilich, S., Kreimer, A., Meilijson, I., Gophna, U., Sharan, R., & Ruppin, E. (2010). The large-scale organization of the bacterial network of ecological co-occurrence interactions. *Nucleic Acids Research, 38*(12), 3857-3868. doi: 10.1093/nar/gkq118

Garrido, M., Rius, F., & Larrechi, M. (2008). Multivariate curve resolution–alternating least squares (MCR-ALS) applied to spectroscopic data from monitoring chemical reactions processes. *Analytical and Bioanalytical Chemistry, 390*(8), 2059-2066. doi: 10.1007/s00216-008-1955-6

Hanauer, S. B. (2006). Inflammatory bowel disease: Epidemiology, pathogenesis, and therapeutic opportunities. *Inflammatory Bowel Diseases, 12*(5), S3-S9. doi: 10.1097/01.mib.0000195385.19268.68

Hendrickson, B. A., Gokhale, R., & Cho, J. H. (2002). Clinical Aspects and Pathophysiology of Inflammatory Bowel Disease. *Clin. Microbiol. Rev., 15*(1), 79-94. doi: 10.1128/cmr.15.1.79-94.2002

Joossens, M., Huys, G., Cnockaert, M., De Preter, V., Verbeke, K., Rutgeerts, P., . . . Vermeire, S. (2011). Dysbiosis of the faecal microbiota in patients with Crohn's

disease and their unaffected relatives. *Gut, 60*(5), 631-637. doi: 10.1136/gut.2010.223263

Karlinger, K., Györke, T., Makö, E., Mester, Á., & Tarján, Z. (2000). The epidemiology and the pathogenesis of inflammatory bowel disease. *European Journal of Radiology, 35*(3), 154-167. doi: 10.1016/s0720-048x(00)00238-2

Kovatcheva-Datchary, P., Zoetendal, E. G., Venema, K., de Vos, W. M., & Smidt, H. (2009). Review: Tools for the tract: understanding the functionality of the gastrointestinal tract. *Therapeutic Advances in Gastroenterology, 2*(4 suppl), s9-s22. doi: 10.1177/1756283x09337646

Kuehbacher, T., Rehman, A., Lepage, P., Hellmig, S., Fölsch, U. R., Schreiber, S., & Ott, S. J. (2008). Intestinal TM7 bacterial phylogenies in active inflammatory bowel disease. *Journal of Medical Microbiology, 57*(12), 1569-1576. doi: 10.1099/jmm.0.47719-0

Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., & Hugenholtz, P. (2008). A Bioinformatician's Guide to Metagenomics. *Microbiol. Mol. Biol. Rev., 72*(4), 557-578. doi: 10.1128/mmbr.00009-08

Lepage, P., Seksik, P., Sutren, M., de la Cochetière, M.-F., Jian, R., Marteau, P., & Doré, J. (2005). Biodiversity of the mucosa-associated microbiota is stable along the distal digestive tract in healthy individuals and patients with IBD. *Inflammatory Bowel Diseases, 11*(5), 473-480. doi: 10.1097/01.mib.0000159662.62651.06

Leser, T. D., Amenuvor, J. Z., Jensen, T. K., Lindecrona, R. H., Boye, M., & Moller, K. (2002). Culture-Independent Analysis of Gut Bacteria: the Pig Gastrointestinal Tract Microbiota Revisited. *Appl. Environ. Microbiol., 68*(2), 673-690. doi: 10.1128/aem.68.2.673-690.2002

Ley, R. E., Peterson, D. A., & Gordon, J. I. (2006). Ecological and Evolutionary Forces Shaping Microbial Diversity in the Human Intestine. *Cell, 124*(4), 837-848. doi: 10.1016/j.cell.2006.02.017

Li, M., Gong, J., Cottrill, M., Yu, H., de Lange, C., Burton, J., & Topp, E. (2003). Evaluation of QIAamp® DNA Stool Mini Kit for ecological studies of gut microbiota. *Journal of Microbiological Methods, 54*(1), 13-20. doi: 10.1016/s0167-7012(02)00260-9

Mai, V., & Draganov, P. V. (2009). Recent advances and remaining gaps in our knowledge of associations between gut microbiota and human health. *World journal of gastroenterology : WJG, 15*(1), 81-85.

Manichanh, C., Rigottier-Gois, L., Bonnaud, E., Gloux, K., Pelletier, E., Frangeul, L., . . . Dore, J. (2006). Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut, 55*(2), 205-211. doi: 10.1136/gut.2005.073817

McOrist, A. L., Jackson, M., & Bird, A. R. (2002). A comparison of five methods for extraction of bacterial DNA from human faecal samples. *Journal of Microbiological Methods, 50*(2), 131-139. doi: 10.1016/s0167-7012(02)00018-0

Mondot, S., Kang, S., Furet, J. P., Aguirre de Carcer, D., McSweeney, C., Morrison, M., . . . Leclerc, M. (2011). Highlighting new phylogenetic specificities of Crohn's disease microbiota. *Inflammatory Bowel Diseases, 17*(1), 185-192. doi: 10.1002/ibd.21436

Nocker, A., Burr, M., & Camper, A. (2007). Genotypic Microbial Community Profiling: A Critical Technical Review. *Microbial Ecology, 54*(2), 276-289. doi: 10.1007/s00248-006-9199-5

Peck, D., Crawford, E., Ross, K., Stegmaier, K., Golub, T., & Lamb, J. (2006). A method for high-throughput gene expression signature analysis. *Genome Biology, 7*(7), R61.

Peterson, D. A., Frank, D. N., Pace, N. R., & Gordon, J. I. (2008). Metagenomic Approaches for Defining the Pathogenesis of Inflammatory Bowel Diseases. *Cell Host & Microbe, 3*(6), 417-427. doi: 10.1016/j.chom.2008.05.001

Picard, C., Fioramonti, J., Francois, A., Robinson, T., Neant, F., & Matuchansky, C. (2005). Review article: bifidobacteria as probiotic agents – physiological effects and clinical benefits. *Alimentary Pharmacology & Therapeutics, 22*(6), 495-512. doi: 10.1111/j.1365-2036.2005.02615.x

Prosser, J. I., Bohannan, B. J. M., Curtis, T. P., Ellis, R. J., Firestone, M. K., Freckleton, R. P., . . . Young, J. P. W. (2007). The role of ecological theory in microbial ecology. [10.1038/nrmicro1643]. *Nat Rev Micro, 5*(5), 384-392.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., . . . Wang, J. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. [10.1038/nature08821]. *Nature, 464*(7285), 59-65. doi: http://www.nature.com/nature/journal/v464/n7285/suppinfo/nature08821_S1.html

Reiff, C., & Kelly, D. (2010). Inflammatory bowel disease, gut bacteria and probiotic therapy. *International Journal of Medical Microbiology, 300*(1), 25-33. doi: 10.1016/j.ijmm.2009.08.004

Round, J. L., & Mazmanian, S. K. (2009). The gut microbiota shapes intestinal immune responses during health and disease. [10.1038/nri2515]. *Nat Rev Immunol, 9*(5), 313-323.

Rudi, K., Zimonja, M., Kvenshagen, B., Rugtveit, J., Midtvedt, T., & Eggesbo, M. (2007). Alignment-Independent Comparisons of Human Gastrointestinal Tract Microbial Communities in a Multidimensional 16S rRNA Gene Evolutionary Space. *Appl. Environ. Microbiol., 73*(8), 2727-2734. doi: 10.1128/aem.01205-06

Rudi, K., Zimonja, M., Trosvik, P., & Næs, T. (2007). Use of multivariate statistics for 16S rRNA gene analysis of microbial communities. *International Journal of Food Microbiology, 120*(1-2), 95-99. doi: 10.1016/j.ijfoodmicro.2007.06.004

Saarela, M., Lähteenmäki, L., Crittenden, R., Salminen, S., & Mattila-Sandholm, T. (2002). Gut bacteria and health foods--the European perspective. *International Journal of Food Microbiology, 78*(1-2), 99-117. doi: 10.1016/s0168-1605(02)00235-0

Salzman, N. H., & Bevins, C. L. (2008). Negative Interactions with the Microbiota: IBD. In G. B. Huffnagle & M. C. Noverr (Eds.), *GI Microbiota and Regulation of the Immune System* (Vol. 635, pp. 67-78): Springer New York.

Sartor, R. B. (1997). Enteric microflora in IBD: Pathogens or commensals? *Inflammatory Bowel Diseases, 3*(3), 230-235. doi: 10.1002/ibd.3780030309

Sartor, R. B. (2006). Mechanisms of Disease: pathogenesis of Crohn's disease and ulcerative colitis. [10.1038/ncpgasthep0528]. *Nat Clin Pract Gastroenterol Hepatol, 3*(7), 390-407.

Shanahan, F. (2009). Therapeutic implications of manipulating and mining the microbiota. *The Journal of Physiology, 587*(17), 4175-4179. doi: 10.1113/jphysiol.2009.174649

Sherry A, D. (2006). Applications of Luminex® xMAP™ technology for rapid, high-throughput multiplexed nucleic acid detection. *Clinica Chimica Acta, 363*(1-2), 71-82. doi: 10.1016/j.cccn.2005.06.023

Skånseng, B., Kaldhusdal, M., & Rudi, K. (2006). Comparison of chicken gut colonisation by the pathogens Campylobacter jejuni and Clostridium perfringens by real-time quantitative PCR. *Molecular and Cellular Probes, 20*(5), 269-279. doi: 10.1016/j.mcp.2006.02.001

Sokol, H., Lay, C., Seksik, P., & Tannock, G. W. (2008). Analysis of bacterial bowel communities of IBD patients: What has it revealed? *Inflammatory Bowel Diseases, 14*(6), 858-867. doi: 10.1002/ibd.20392

Sokol, H., Pigneur, B., Watterlot, L., Lakhdari, O., Bermúdez-Humarán, L. G., Gratadoux, J.-J., . . . Langella, P. (2008). Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients.

Proceedings of the National Academy of Sciences, 105(43), 16731-16736. doi: 10.1073/pnas.0804812105

Stange, E. F., Travis, S. P. L., Vermeire, S., Beglinger, C., Kupcinkas, L., Geboes, K., . . . Reinisch, W. (2006). European evidence based consensus on the diagnosis and management of Crohn's disease: definitions and diagnosis. *Gut, 55*(suppl 1), i1-i15. doi: 10.1136/gut.2005.081950a

Stange, E. F., Travis, S. P. L., Vermeire, S., Reinisch, W., Geboes, K., Barakauskiene, A., . . . Warren, B. F. (2008). European evidence-based Consensus on the diagnosis and management of ulcerative colitis: Definitions and diagnosis. *2*(1), 1-23.

Swerdlow, H., & Gesteland, R. (1990). Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Research, 18*(6), 1415-1419. doi: 10.1093/nar/18.6.1415

Trosvik, P., Skanseng, B., Jakobsen, K. S., Stenseth, N. C., Naes, T., & Rudi, K. (2007). Multivariate Analysis of Complex DNA Sequence Electropherograms for High-Throughput Quantitative Analysis of Mixed Microbial Populations. *Appl. Environ. Microbiol., 73*(15), 4975-4983. doi: 10.1128/aem.00128-07

Tuohy, K. M., Gougoulias, C., Qing, S., Walton, G., Fava, F., & Ramnani, P. (2009). Studying the Human Gut Microbiota in the Trans-Omics Era - Focus on Metagenomics and Metabonomics. [Article]. *Current Pharmaceutical Design, 15*(13), 1415-1427.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The Human Microbiome Project. [10.1038/nature06244]. *Nature, 449*(7164), 804-810.

Vebo, H. C., Sekelja, M., Nestestog, R., Storro, O., Johnsen, R., Oien, T., & Rudi, K. (2011). Temporal development of the infant gut microbiota in IgE sensitized and non-sensitized children determined by the GA-map infant array. *Clin. Vaccine Immunol.*, CVI.00062-00011. doi: 10.1128/cvi.00062-11

Ventura, M., O'Flaherty, S., Claesson, M. J., Turroni, F., Klaenhammer, T. R., van Sinderen, D., & O'Toole, P. W. (2009). Genome-scale analyses of health-promoting bacteria: probiogenomics. [10.1038/nrmicro2047]. *Nat Rev Micro, 7*(1), 61-71.

Woese, C. R. (1987). Bacterial evolution. *Microbiological Reviews, 51*(2), 221-271.

Xavier, R. J., & Podolsky, D. K. (2007). Unravelling the pathogenesis of inflammatory bowel disease. [10.1038/nature06005]. *Nature, 448*(7152), 427-434.

Appendix

Table 2.1.1A: Samples and Diagnosis.

Sample ID	Diagnosis	Sample ID	Diagnosis	Sample ID	Diagnosis
7	CD	180	CD	6006	CONTROL
20	CONTROL	181	CD	6007	UC
28	CD	187	CONTROL	6008	CONTROL
32	CONTROL	188	CD	6009	CONTROL
35	CONTROL	191	CONTROL	6010	CD
50	CONTROL	192	UC	6011	UC
54	CONTROL	195	CONTROL	6013	CD
58	CONTROL	196	CONTROL	6014	IBDU
71	CD	198	UC	6015	CONTROL
72	CONTROL	201	CONTROL	6016	CONTROL
77	CONTROL	202	CD	6017	UC
79	UC	205	CONTROL	6018	CD
86	UC	207	CONTROL	6019	CD
87	CONTROL	1005	CD	6021	CONTROL
88	CONTROL	1006	CONTROL	6022	CD
99	CONTROL	1013	UC	6023	UC
111	CD	1028	CD	6026	UC
119	UC	1030	CD	6028	CD
126	UC	2024	CD	6030	CD
130	CONTROL	2025	IBDU	6032	CD
131	CD	5009	CONTROL	6033	UC

136	CONTROL	6001	UC	6034	CD
158	CONTROL	6003	CD	6035	CD
163	CD	6004	CONTROL	6036	CD
172	CD	6005	UC	6038	UC

Table 2.1.2A: Diagnosis and classification criteria for IBD as either CD, UC or IBDU.

Classification	**Diagnosis features**
CD	• Clinical features including abdominal pain, diarrhoea and weight loss • Macroscopic appearance at operation or endoscopy: segmental, discontinuous, and/or patchy lesions with or without rectal involvement, discrete or apthous ulcerations, fissuring or penetrating lesions, cobblestone or strictures • Radiological evidence of stenosis in the small bowel, segmental colitis or findings of fistulae • Histologic evidence of transmural inflammation or epithelial granulomas with giant cells.
UC	• A history of diarrhoea and or blood/pus in stool • Macroscopic appearance at endoscopy, with continuous mucosal inflammation affecting the rectum in continuity with some or the entire colon • Microscopic features on biopsy compatible with UC • No suspicion of CD on small bowel ileocolonoscopy or biopsy.
❖ Patients with inconclusive or divergent endoscopy and histopathology according to CD or UC criteria were classified as IBDU.	

Chimeric sequence removal with chimera slayer algorithm in mothur :

1. Sequences in fasta format
2. Mothur (program)
3. >summary.seqs (fasta=….)
4. May do: Filtrate sequences (n, homopolymers, primers, seq that don`t have the expected length)
5. >align.seqs(candidate=….., template=……, ksize=8)

template file: rRNA16S.gold.NAST_ALIGNED.fasta

6. >filter.seqs(fasta=…..align-template.fasta, vertical=T)
7. >chimera.slayer(fasta=…filter.fasta, template=aligned.filter.fasta, minbs=95) chimera removal base don the chimara slayer algorithm
8. >remove.seqs(accnos=…accnos, fasta=…unique.fasta, names=….names)
9. >summary.seqs(fasta=unique.pick.fasta)
10. create names file: >unique.seqs(fasta=…..fasta) identify the unique sequences in a collection and generate a names file

Figure 2.8.1A: Mothur commands (Chimeric Slayer).

Multivariate Curve Resolution steps for bacterial mixed sequences: Commands used.

The large set of mixed sequences -> aligned all of them -> took the aligned spectra and pre-processed/normalized them (aligned once again, but taking only small portions (individual peaks) at a time to ensure that there were no small peak shifts due to differences in the retention time) ->decided how much information that could be got out of the dataset (perform PCA and/or EFA [Evolving Factor Analysis] to decide. Result - number of 'significant' components which explain the most of variation) -> took these number of components, perform MCR (just upload the file with aligned spectra to the program, set the number of components to pre-determined value, click 'run') -> get two outputs. One contained the information on the relative amount of each of the components in every sample/individual of the dataset. The other had the spectral information on what each of the components are -> plotted the spectral information and basecalled -> BLASTed the sequence to identify what each of the components were.

Figure 2.9.1A: Commands operations used for MCR of mixed bacterial sequences.

Table 3.2.1A: Quality of Sequences, Illustration of the Sequence quality based on the sequencer capillary signal. Saturated ≥8000; Good ≥2000; Moderate ≥1000 or <2000; Poor ≤1000

Sample Number	Sequence Quality	Sample Number	Sequence Quality	Sample Number	Sequence Quality
7	Good	180	Good	6007	Moderate
20	Good	181	Good	6008	Moderate
28	Good	187	Good	6009	Poor
32	Good	188	Moderate	6010	Good
35	Good	191	Good	6011	Moderate
50	Good	192	Poor	6013	Moderate
54	Poor	195	Poor	6014	Good
58	Good	196	Moderate	6015	Good
71	Good	198	Good	6016	Moderate
72	Noise	201	Good	6017	Good
77	Saturated	202	Noise	6018	Good
79	Good	205	Moderate	6019	Good
86	Good	207	Poor	6021	Good

87	Good	1005	Moderate	6022	Good
88	Good	1006	Good	6023	Good
99	Good	1013	Good	6026	Good
111	Moderate	1028	Good	6028	Good
119	Good	1030	Moderate	6030	Good
126	Good	2025	Moderate	6032	Moderate
130	Good	5009	Poor	6033	Good
131	Moderate	6001	Moderate	6034	Good
136	Good	6003	Good	6035	Good
158	Poor	6004	Moderate	6036	Good
163	Good	6005	Moderate	6038	Good
172	Moderate	6006	Moderate	6039	Noise

Table 3.1.1A: DNA quality measure by Nanodrop spectrophotometry. For pure DNA, $A_{260/280}$ is ~1.8 and for pure RNA $A_{260/280}$ is ~2. DNA gel band strength and degradation are graded: Strong (+++), Good (++) and Weak (+). No trend was observed for the varying DNA band signals.

Sample #	Gel band Evaluation		OD Evaluation		Sample #	Gel band Evaluation		OD Evaluation		Sample #	Gel band Evaluation		OD Evaluation	
	Strength	degradation	260/280	quality		Strength	degradation	260/280	quality		Strength	degradation	260/280	quality
7	+	+	2.15	84.01	180	+	+	1.76	23.65	6007	+	+	2.62	16.11

20	+	+	1.95	43.84	181	++	+	1.32	3.74	6008	++	+	1.83	45.47
28	+	0	1.79	9.5	187	+	0	6.30	2.44	6009	+	+++	2.26	98.6
32	+	0	2.19	3.86	**188**			1.98	10.0	6010	+++	0	2.26	25.73
35	+	0	1.65	5.7	191	++	+++	2.20	118.41	6011	+++	0	2.47	10.84
50	++	+++	2.03	41.42	192	+++	0	2.22	49.58	6013	+	+	1.19	38.95
54	++	+	1.97	42.74	195	+++	0	2.64	10.24	6014	+	++	2.17	98.99
58	++	+++	2.13	149.05	196	++	+++	2.23	111.11	6015	++	+++	2.17	46.87
71	+	+	1.97	20.6	198	+	0	2.65	4.39	6016	+++	+	1.71	15.11
72	+	0	1.99	26.92	201	+++	0	2.77	7.66	6017	+	0	2.17	60.8
77	++	+	1.88	9.45	202	+++	+	2.17	33.71	6018	+	++	1.92	20.86
79	+++	+++	1.86	47.32	**205**			1.99	10.0	6019	+	0	1.79	8.31
86	+++	++	1.59	11.62	207	+	+++	2.23	89.56	6021	+	0	2.24	5.6
87	+++	0	1.84	26.96	1005	++	++	2.21	58.11	6022	+	0	2.21	23.43
88				10.0	1006	+++	+	2.25	39.45	6023	+++	0	2.02	117.56
99	++	+	1.85	65.32	1013	++	+	2.31	35.13	6026	+++	+++	2.10	6.34

111	++	+++	1.96	88.05	**1028**	++		2.35	10.0	6028	+	0	2.09	6.45
119	++	0	2.00	17.96	**1030**			2.51	6.0	6030	+	0	2.10	79.47
126	+	+	2.00	35.3	**2025**	+		2.17	10.0	6032	++	+++	2.20	118.81
130	+	+	1.97	35.19	5009	+	+	2.07	31.35	6033	+	+++	2.15	33.88
131	+	+	2.11	29.94	6001	++	+	2.34	18.08	6034	+	+	2.15	43.29
136	+	+	2.08	41.09	6003	+	+	2.43	15.59	6035	+	+	2.04	123.71
158	+	0	0.71	1.35	6004	+	+	2.36	32.11	6036	++	++	1.83	12.77
163			2.11	10.0	6005	+	+++	2.32	66.59	6038	+	0	2.17	106.68
172			1.87	10.0	6006	+	0	2.71	16.7	6039	+++	+++	2.45	6.87

Table 3.3.2A: Clone library. Sequences match by Ribosomal Database Project.

Disease	**Clone ID**	**Domain**	**Phylum**	**Class**	**Order**	**Family**	**Genus**
CD	111.2	Bacteria	Verrucomicrobia	Verrucomicrobiae	Verrucomicrobiales	Verrucomicrobiaceae	Akkermansia
CD	111.16	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	Alistipes
CD	172.18	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	Alistipes
CON	5009.25	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	Alistipes
CON	5009.5	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Rikenellaceae	Alistipes
CD	111.23	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
CD	111.9	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides

CD	172.12	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
CD	172.31	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
CD	6034.21	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
CD	6034.28	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
CD	6036.35	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
CD	6036.4	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
CON	32.5	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
CON	5009.17	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
UC	198.1	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
UC	198.26	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
UC	198.28	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
UC	198.36	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
UC	198.37	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
UC	198.39	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
UC	6005.28	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
UC	6005.8	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
UC	86.1	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
UC	86.12	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
UC	86.17	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
UC	86.2	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
UC	86.2	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
UC	86.26	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
UC	86.3	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
UC	86.32	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
UC	86.9	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
UC	87.1	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
UC	87.25	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides

UC	87.28	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
UC	87.9	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Bacteroidaceae	Bacteroides
CD	6032.39	Bacteria	Firmicutes	Clostridia	Clostridiales	Incertae Sedis XIV	Blautia
CD	111.37	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Butyricicoccus
CD	6036.21	Bacteria	Firmicutes	Clostridia	Clostridiales	Clostridiaceae (Subfamily: Clostridiaceae 1)	Clostridium
CD	6034.7	Bacteria	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	Dialister
CON	5009.11	Bacteria	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	Dialister
CON	5009.12	Bacteria	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	Dialister
CON	5009.27	Bacteria	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	Dialister
CON	5009.32	Bacteria	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	Dialister
CON	5009.9	Bacteria	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	Dialister
UC	6005.11	Bacteria	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	Dialister
UC	6005.12	Bacteria	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	Dialister
UC	6005.14	Bacteria	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	Dialister
UC	6005.15	Bacteria	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	Dialister
UC	6005.16	Bacteria	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	Dialister
UC	6005.19	Bacteria	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	Dialister
UC	6005.2	Bacteria	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	Dialister
UC	6005.2	Bacteria	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	Dialister
UC	6005.21	Bacteria	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	Dialister
UC	6005.23	Bacteria	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	Dialister
UC	6005.32	Bacteria	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	Dialister
UC	6005.34	Bacteria	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	Dialister
UC	6005.36	Bacteria	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	Dialister
UC	6005.38	Bacteria	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	Dialister
UC	6005.4	Bacteria	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	Dialister

CD	172.1	Bacteria	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Enterococcus
CD	172.21	Bacteria	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Enterococcus
CD	172.23	Bacteria	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Enterococcus
CD	172.25	Bacteria	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Enterococcus
CD	172.26	Bacteria	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Enterococcus
CD	172.27	Bacteria	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Enterococcus
CD	172.29	Bacteria	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Enterococcus
CD	172.3	Bacteria	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Enterococcus
CD	172.33	Bacteria	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Enterococcus
CD	172.34	Bacteria	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Enterococcus
CD	172.35	Bacteria	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Enterococcus
CD	172.38	Bacteria	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Enterococcus
CD	172.39	Bacteria	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Enterococcus
CD	172.6	Bacteria	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Enterococcus
CD	1030.13	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CD	1030.16	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CD	1030.17	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CD	1030.18	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CD	1030.2	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CD	1030.21	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CD	1030.23	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CD	1030.3	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CD	1030.31	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella

CD	1030.33	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CD	1030.34	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CD	1030.37	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CD	1030.7	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CD	1030.8	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CD	172.13	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CD	172.24	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CD	6034.1	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CD	6034.13	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CD	6034.31	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CON	32.11	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CON	32.12	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CON	32.15	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CON	32.16	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CON	32.19	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CON	32.2	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CON	32.21	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CON	32.22	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella

CON	32.23	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CON	32.24	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CON	32.25	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CON	32.28	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CON	32.3	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CON	32.3	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CON	32.35	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CON	32.36	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CON	32.37	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CON	32.4	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CON	32.6	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CON	32.8	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
UC	119.1	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
UC	119.11	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
UC	119.12	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
UC	119.16	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
UC	119.17	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
UC	119.2	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella

UC	119.21	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
UC	119.22	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
UC	119.23	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
UC	119.24	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
UC	119.25	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
UC	119.26	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
UC	119.29	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
UC	119.3	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
UC	119.31	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
UC	119.33	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
UC	119.35	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
UC	119.36	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
UC	119.39	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
UC	119.4	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
UC	119.5	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
UC	119.8	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella
CD	6032.21	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Faecalibacterium
CON	6009.19	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Faecalibacterium
CON	6009.23	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Faecalibacterium

CON	6009.9	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Faecalibacterium
UC	86.14	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Faecalibacterium
UC	86.19	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Faecalibacterium
UC	86.21	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Faecalibacterium
UC	86.25	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Faecalibacterium
UC	86.29	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Faecalibacterium
UC	86.3	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Faecalibacterium
UC	86.38	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Faecalibacterium
UC	86.5	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Faecalibacterium
CD	6036.3	Bacteria	Firmicutes	Bacilli	Lactobacillales	Carnobacteriaceae	Granulicatella
CD	6036.5	Bacteria	Firmicutes	Bacilli	Lactobacillales	Carnobacteriaceae	Granulicatella
CD	6034.14	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
CD	6034.19	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
CD	6034.2	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
CD	6034.22	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
CD	6034.25	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
CD	6036.1	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
CD	6036.19	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
CD	6036.2	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
CD	6036.2	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
CD	6036.23	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
CD	6036.27	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus

CD	6036.29	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
CD	6036.3	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
CD	6036.32	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
CD	6036.33	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
CD	6036.37	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
UC	198.1	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
UC	198.14	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
UC	198.16	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
UC	198.17	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
UC	198.2	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
UC	198.21	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
UC	198.22	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
UC	198.23	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
UC	198.24	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
UC	198.27	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
UC	198.32	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
UC	198.34	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
UC	198.35	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus

UC	198.38	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
UC	198.4	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
UC	198.7	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
UC	198.8	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	Haemophilus
CD	6032.31	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
UC	6001.12	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
UC	6001.15	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
UC	6001.16	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
UC	6001.2	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
UC	6001.22	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
UC	6001.24	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
UC	6001.3	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
UC	6001.3	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
UC	6001.31	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
UC	6001.33	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
UC	6001.37	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
UC	6001.38	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
UC	6001.39	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
UC	6001.4	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
UC	6001.5	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
UC	6001.6	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
UC	6001.7	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
UC	6001.11	Bacteria	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
CD	6032.25	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Lactococcus
CD	6032.6	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Lactococcus

CON	5009.34	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Lactonifactor
CON	32.33	Bacteria	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	Megamonas
UC	198.6	Bacteria	Proteobacteria	Betaproteobacteria	Neisseriales	Neisseriaceae	Neisseria
UC	6005.24	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	Odoribacter
CD	172.2	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Oscillibacter
CON	5009.29	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Oscillibacter
CON	5009.33	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Oscillibacter
CD	111.3	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	Parabacteroides
CD	111.32	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	Parabacteroides
CON	6009.6	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Porphyromonadaceae	Parabacteroides
CON	32.26	Bacteria	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	Phascolarctobacterium
UC	6005.37	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella
UC	6005.7	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella
CON	6009.2	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Roseburia
CON	6009.4	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Roseburia
CON	6009.5	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	Roseburia
CD	6032.19	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcus
CD	6032.26	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcus
CD	6032.28	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcus
CD	6032.4	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus
CD	6036.24	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus
CD	6036.31	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus
CD	6036.6	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus
CD	6032.11	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Subdoligranulum
CD	6032.12	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Subdoligranulum
CD	6032.14	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Subdoligranulum

CD	6032.17	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Subdoligranulum
CD	6032.18	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Subdoligranulum
CD	6032.22	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Subdoligranulum
CD	6032.23	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Subdoligranulum
CD	6032.27	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Subdoligranulum
CD	6032.29	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Subdoligranulum
CD	6032.3	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Subdoligranulum
CD	6032.32	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Subdoligranulum
CD	6032.35	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Subdoligranulum
CD	6032.37	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Subdoligranulum
CD	6032.7	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Subdoligranulum
CD	6032.9	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Subdoligranulum
CD	6032.15	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Subdoligranulum
CD	6034.26	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Subdoligranulum
CON	6009.1	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Subdoligranulum
CON	6009.7	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Subdoligranulum
UC	6005.17	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Subdoligranulum
UC	6005.29	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Subdoligranulum
UC	6005.35	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	Subdoligranulum
CD	172.19	Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Alcaligenaceae	Sutterella
CD	6036.1	Bacteria	Firmicutes	Clostridia	Clostridiales	Peptostreptococcaceae	unclassified_"Peptostreptococcaceae
CD	6032.13	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	unclassified_Lachnospiraceae
CD	6034.24	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	unclassified_Lachnospiraceae
CON	5009.14	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	unclassified_Lachnospiraceae

CON	5009.2	Bacteria	Firmicutes	Clostridia	Clostridiales	Lachnospiraceae	unclassified_Lachnospiraceae
CD	6034.3	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	unclassified_Pasteurellaceae
CD	6034.6	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	unclassified_Pasteurellaceae
CD	6036.13	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	unclassified_Pasteurellaceae
UC	198.11	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	unclassified_Pasteurellaceae
UC	198.12	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	unclassified_Pasteurellaceae
UC	198.15	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	unclassified_Pasteurellaceae
UC	198.3	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	unclassified_Pasteurellaceae
UC	198.31	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	unclassified_Pasteurellaceae
UC	198.33	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	unclassified_Pasteurellaceae
UC	198.9	Bacteria	Proteobacteria	Gammaproteobacteria	Pasteurellales	Pasteurellaceae	unclassified_Pasteurellaceae
CON	6009.28	Bacteria	Firmicutes	Clostridia	Clostridiales	Ruminococcaceae	unclassified_Ruminococcaceae
UC	86.36	Bacteria	Firmicutes	Clostridia	Clostridiales	Veillonellaceae	unclassified_Veillonellaceae
CD	111.4	Bacteria	Firmicutes	unclassified_"Firmicutes			
CD	6032.15	Bacteria	Firmicutes	Clostridia	Clostridiales	unclassified_Clostridiales	
CD	6032.33	Bacteria	Firmicutes	Clostridia	Clostridiales	unclassified_Clostridiales	
CD	6032.34	Bacteria	Firmicutes	Clostridia	Clostridiales	unclassified_Clostridiales	
CD	6034.32	Bacteria	Firmicutes	Clostridia	Clostridiales	unclassified_Clostridiales	
CD							

CD	6036.28	Bacteria	Firmicutes	unclassified_Firmicutes			
CD	6036.7	Bacteria	TM7	TM7_genera_incertae_sedis			
CON	32.1	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	unclassified_Bacteroidales	
CON	5009.2	Bacteria	Firmicutes	Clostridia	Clostridiales	unclassified_Clostridiales	
CON	6009.22	Bacteria	Firmicutes	Clostridia	Clostridiales	unclassified_Clostridiales	
UC	6005.25	Bacteria	Firmicutes	Clostridia	Clostridiales	unclassified_Clostridiales	
UC	6005.3	Bacteria	Firmicutes	Clostridia	Clostridiales	unclassified_Clostridiales	
UC	86.24	Bacteria	Firmicutes	Clostridia	Clostridiales	unclassified_Clostridiales	
UC	86.28	Bacteria	Firmicutes	Clostridia	Clostridiales	unclassified_Clostridiales	
UC	86.39	Bacteria	Firmicutes	Clostridia	Clostridiales	unclassified_Clostridiales	
UC	86.4	Bacteria	Firmicutes	Clostridia	Clostridiales	unclassified_Clostridiales	

Table 3.3.3A: MCR data showing average component frequency for each sample. Control=1,CD=2, UC=3 and IBDU=4

Sample ID	Disease Diagnostic	av comp1	av comp3	av comp4	av comp5	av comp6
1005	2	0.09662275	0.10094035	0.1778055	1.15541	0.266845
1006	1	0.025005	0.01839855	0.0144962	1.73841	0.03794785
1013	3	0.8957145	0.00305693	0.0493003	0.11463055	0.592095
1028	2	0.0168919	0.0345116	0.02519813	1.71786	0.01891755
1030	2	0.00740699	0.02378553	0.01942797	1.76478	0.0091893
111	2	0.51816133	0.0831854	0.08442183	0.0861447	0.83792167
119	3	0.0386527	0.02281594	0.03844705	1.70626	0
126	3	0.62727533	0.30694733	0.198882	0.17261667	0.302465
130	1	0.304723	0.0714056	0.104621	0.345518	0.867447
131	2	0.292211	0.217679	0.593322	0.234191	0.359489
136	1	0.27838	0.17847967	0.643364	0	0.49705033
158	1	0.2777785	0.00615805	0.162374	0	1.15297
163	2	0.40003633	0.10785117	0.07688933	0.786704	0.306434

172	2	0.2502545	0.43028	0.03650655	0.455322	0.583656
180	2	0.3135775	0.0676525	0	1.10081	0.29715
181	2	0.4978575	0.02679835	0.0911631	0.67396425	0.434388
187	1	0.182942	0	0.124803	1.12592	0.355492
188	2	0.5983775	0.05192515	0.07374915	0.649892	0.2713805
191	1	0.02105155	0.01310901	0.00775747	1.699615	0.08128055
192	3	0.657207	0.0881509	0.140398	0.047746	0.635054
195	1	0.5114955	0.00440618	0.08664225	0.0509506	0.921095
196	1	0.46484233	0.159627	0.0843206	0.737546	0.231428
198	3	0	0	1.295265	0.485519	0.03824685
20	1	0.43667967	0.0170419	0.0975704	1.062747	0.13136883
201	1	0.13187697	0	0.12564367	1.11919667	0.36285867
202	2	0.127601	0.130105	0.134221	1.33994	0
2025	4	0.93864133	0.10505987	0.0945611	0.0721556	0.347344
205	1	0.5822895	0.04801625	0.10945465	0.41525055	0.4709825
207	1	0.311019	0.20256	0.546623	0.374602	0.20931
28	2	0.6704415	0.160596	0.0613596	0.370275	0.37944
32	1	0.04704057	0.03782793	0.03812417	1.67478	0.01684512
35	1	0.17998733	0.0084628	0.0369488	1.36074	0.20583367
50	1	0.79215667	0.419869	0.24483633	0	0.08123027
5009	1	0.14277633	1.39399667	0.03259643	0.02722243	0
54	1	0.81685967	0.10989567	0.11237667	0.191019	0.350344
58	1	0.4348495	0.02663475	0.1381545	0.00723789	1.0045865
6001	3	0	0.5696545	0	0.2945825	0.911681
6003	2	0.672149	0.5088505	0.2700145	0	0.06372525
6004	1	0.413348	0.1166487	0.15985207	0.43103033	0.50239167
6005	3	0.06622063	1.39467667	0.00116812	0.10678643	0.04393467
6006	1	0	0.543341	0	0.4984615	0.722222
6007	3	0.030076	0.0586965	0.0003734	0	0
6008	1	0.372814	0.26736	0.301091	0.6086495	0.0352994
6009	1	1.30767333	0	0	0.17087133	0.154097
6010	2	0.48467133	0.0250996	0	0.921392	0.27751867
6011	3	0.51100233	0.220211	0.22335467	0.13185913	0.50389333
6013	2	0.231449	0.12324673	0.65559333	0.67316167	0.010138
6014	4	0.155117	0.0368242	0.02107873	1.510605	0.09391367
6015	1	0.609845	0.052221	0.00850015	0.7537705	0.2387575
6016	1	0.6064225	0.0866431	0.2063235	0.294706	0.444143
6017	3	1.197365	0	0	0.3812485	0.10411235
6018	2	0.810702	0.04685035	0.1048536	0.08954665	0.5524885
6019	2	0.682762	0.0533344	0.00886335	0.2555735	0.644604
6021	1	1.184275	0.0662752	0.0050822	0.145912	0.20236
6022	2	0.07606355	0.06641105	0.09182045	1.365915	0.1597211
6023	3	0.9823295	0.06487415	0.01284244	0.1596155	0.3957065
6026	3	0.01129195	0.01243805	0.0118152	1.75807	0
6028	2	0.222192	0.0965146	0.052081	1.34487	0.0712833

6030	2	1.0196655	0.04346545	0.1774935	0.03632865	0.3212835
6032	2	1.18717	0.07119985	0.0102897	0.241384	0.1453905
6033	3	0.07865965	0	0.2021825	1.510615	0
6034	2	0.3155155	0.0788566	0.7212625	0.6128485	0
6035	2	0.7425085	0.00256881	0.1443815	0	0.693175
6036	2	0.2415125	0.0694424	0.945125	0.3673285	0.06244665
6038	3	0.738363	0.2405225	0.191375	0.04415315	0.40472
7	2	0.0692445	0.02751875	0.0094975	1.672575	0.04123755
71	2	0.172603	0.03256627	0.094674	1.50565	0.00268708
77	1	0.05496753	0.0569889	0.0459642	1.62066333	0.0136956
79	3	0.296495	0.141944	0.2653185	0.937086	0
86	3	1.20533	0.0114154	0	0.238771	0.151405
87	1	0.1975235	0	0.129847	0.0168531	1.31595
88	1	0.610152	0.0571016	0.136338	0.0785317	0.718745
99	1	0.773552	0.0195371	0.0882845	0.0866485	0.661091

Table 3.4.1A: List of the 19 probes designed and evaluated with targets and non-targets from the clone library.

Probe Number	Sequence	Genus (Target)
Probe_1	CCTCCAGTTTATCACTGGC	E.Coli
Probe_2	TCAAGCTTGCCAGTATCAGATGC	E.Coli
Probe_3	GCCTCAAGGGCACAAC	E.Coli
Probe_4	CGCGGGTTCGCTTCTCTTTGTTGAC	Dialister
Probe_5	GACATTGATCGCGATCTGCAGAAATGC	Dialister
Probe_6	AAGAACTCCGCATTTCTGC	Dialister
Probe_7	ACAACGCTTGTGACCTAC	Faecalibacterium
Probe_8	CGTAGTTAGCCGTCACTTC	Faecalibacterium
Probe_9	CGGACAACGCTTGTGAC	Faecalibacterium
Probe_10	CCCAGTATGAAATGCAATTCCC	Haemophilus
Probe_11	ACCGCAACATTCTGATTTGC	Haemophilus
Probe_12	CGCTTCCCTCTGTATACGC	Haemophilus
Probe_13	TCGCTTCCCTCTGTATACG	Haemophilus
Probe_14	CGATTAACGCTCGCAC	Haemophilus
Probe_15	CCTCCAACACTTAGCACT	Enterococcus
Probe_16	CCCTCCAACACTTAGCA	Enterococcus
Probe_17	CCTGTCACTTTGCCC	Enterococcus
Probe_18	CCTGTTTGCTACCCATACTTT	Lactobacillus
Probe_19	GAGTTCCACTGTCCTCTT	Lactobacillus
Probe_20	CCATGCACCACCTGTAT	Lactobacillus
Probe_21	CTTGCTTCTCTTTGTTTAACGC	Faecalibacterium
Probe_22	GCTTCTCTTTGTTTAACGCC	Faecalibacterium
Probe_23	GGTCTTGCTTCTCTTTGTTTAAC	Faecalibacterium
Probe_U (UNI01)	CGTATTACCGCGGCTGCTGGCA	Universal

Table 3.4.2A: Probe Data showing normalised “peak height” (PH)

Sample Number	NP_3H	NP_6H	NP_8H	NP_13H	NP_16H	NP_18H	P_Universal
7	0.965185	51.82609	216.7273	59.10744	143.04	3.445087	52472
28	0.398071	35.64063	6.912121	50.68889	57.025	2.04574	18447

71	0.528562	0	11.44094	28.77228	0	2.310016	23602
79	0.391951	106.7059	17.87192	0	14.11673	4.365824	26210
86	0.096728	7.478723	1.604412	60.25714	114	1.151515	30104
111	0.013506	29.8498	22.54328	62.41322	92.09756	3.186498	57739
119	1.313894	135.75	67.875	0	72.9403	4.49586	32488
126	0.108243	4.167482	11.21382	20.1716	85.225	2.751412	25054
131	0.018779	23.02703	2.843159	4.689908	42.6	5.496774	20079
163	0.391379	6.286022	7.066076	48.98883	153.8421	2.798022	68611
172	0.103516	0	5.818182	0	0.70137	2.064516	3957
180	0.683317	37.77949	0	31.89177	237.6452	3.906151	55413
181	0.013017	10.95796	3.426291	23.61812	105.7681	2.407786	53922
188	0.207226	0	4.079533	122.25	0	2.324144	54459
192	0.016413	0	9.102508	25.76235	0	2.57544	63641
198	0.005523	0	115.7705	2.337637	108.6462	11.80936	49993
202	0.839827	0	3.652611	0	0	3.951788	56663
1005	0.446633	54.9881	14.52516	50.75824	124.8378	7.711185	30982
1013	0.005176	143.9412	8.220605	92.92405	103.3944	2.150894	54072
1028	1.020326	171.5641	32.79902	0	89.21333	6.144169	44117
1030	1.164364	136.6889	72.36471	46.24812	74.10843	5.42895	39406
2025	0.005126	12.6826	4.307143	90.86301	157.9286	1.594855	43103
6001	0.010704	141.3333	16.16422	78.74286	2.199521	0.656112	35776
6003	0.04615	3.104255	14.11935	59.14865	61.64789	3.276198	29416
6005	0.01027	64.91111	2.507296	36.06173	78.94595	1.947333	21571
6007	1.42129	110.85	33.84733	134.3636	94.34043	3.359091	29601
6010	0.524477	96.52941	19.61355	54.0989	87.91071	2.283395	33060
6011	0.117002	5.372159	10.6087	88.98824	0	4.339644	56321
6013	0.435301	0	40.43452	2.975471	52.65891	6.11982	45041
6014	0.902774	13.20513	11.28326	200.2778	0	3.221626	50639
6017	0.168671	35.77907	4.478894	45.92537	96.15625	0.724682	22810
6018	0.046493	49.9	8.754386	69.30556	97.84314	3.015106	32726
6019	0.076756	75.56364	24.7381	30.11594	62.02985	2.581366	27769
6022	0.71022	15.03546	76.62651	27.06383	94.92537	4.937888	41219
6023	0.006026	20.15835	5.554692	61.95333	127.3014	1.278619	70161
6026	1.174614	122.0492	91.91358	0	225.6061	6.209341	53762
6028	0.685225	85.14815	42.57407	35.55155	54.7381	6.315934	45776
6030	0.029923	39.84615	5.853107	24.37647	45.04348	1.248193	15402
6032	0.096943	18.46774	8.178571	14.87013	9.956522	0.555556	8583
6033	2.263868	39.23529	46.80702	7.847059	47.64286	6.262911	20664
6034	0.353659	0	18.65868	2.670094	89.02857	1.859189	23377
6035	0.004894	18.43609	7.710692	67.48624	156.5106	2.244051	53849
6036	0.227696	8.344444	3.413636	0.97406	8.075269	1.203526	5694
6038	0.017297	3.567985	4.257768	46.83544	63.7931	2.306733	24914
6039	0	0	1.882353	0	0	1.103448	541
20	0.29715	0	51.57576	130.9231	0	1.295775	48858
32	1.001237	0	45.75472	0	0	3.479197	19583

35	0.602475	35.21795	6.782716	67	76.30556	2.051531	21773
50	0	0.270887	2.222615	4.278912	0	0.161158	6329
54	0.081458	0	5.463547	57.71264	156.9063	2.725841	34611
58	0	0	0.211538	1.5	0	0.062937	1015
72	0	0	0	0	0	0	212
77	1.010471	5.711522	88.9589	216.4667	0	4.733236	44387
87	0.008693	71.10909	10.21149	30.79528	55.87143	3.055469	28488
88	0	74.2	7.756098	25.29545	0	0	17873
99	0	165.5455	32.37333	148.6531	108.7164	2.298517	57741
130	0.110646	50.57692	26.56566	46.14035	16.3354	3.902077	19681
136	0.39404	0.606426	2.323077	0.816216	7.74359	0.444118	2874
158	0	52.78873	22.17751	0	0	4.322953	26525
187	0	0	47.91509	44.9469	0	5.732506	33808
191	1.043839	0	63.12791	92.01695	169.6563	4.600847	35260
195	0.029853	0	6.745501	74.97143	0	2.707946	60707
196	0.240115	0	32.34071	112.4462	235.7742	2.321054	51918
201	0.683189	0	29.55814	206.1081	217.8857	4.796226	57156
205	0.011197	99.45455	23.15344	61.6338	125.0286	4.151803	29390
207	0.302995	4.96332	15.87037	3.720695	71.41667	2.081781	19669
1006	1.074778	8.965909	12.32813	9.740741	11.60294	2.72069	5991
5009	0	1.684906	34.23881	93.63265	148	3.13388	30273
6004	0.266638	4.278628	11.0205	137	134.3654	2.471525	49354
6006	0.160913	87.5814	63.83051	0	4.922876	2.850871	25528
6008	0.434013	0	16.95143	44.94697	15.94892	5.962814	37291
6009	0.030303	57.2	1.625	33	39.44828	0.594492	24480
6015	0.341272	90.44231	16.38676	85.50909	142.5152	1.547039	30215
6016	0.210776	9.032086	3.526096	34.46939	42.225	1.864238	12522
6021	0.005796	137.3889	2.296193	95.11538	200.5135	1.019934	53929